Psychological Injury and Law
https://doi.org/10.1007/s12207-020-09389-4

The Development of a Short Version of the SIMS Using Machine Learning to Detect Feigning in Forensic Assessment

Graziella Orrù[1] · Cristina Mazza[2] · Merylin Monaro[3] · Stefano Ferracuti[4] · Giuseppe Sartori[3] · Paolo Roma[4]

Received: 4 May 2020 / Accepted: 14 September 2020

Abstract
In the present study, we applied machine learning techniques to evaluate whether the Structured Inventory of Malingered Symptomatology (SIMS) can be reduced in length yet maintain accurate discrimination between consistent participants (i.e., presumed truth tellers) and symptom producers. We applied machine learning item selection techniques on data from Mazza et al. (2019c) to identify the minimum number of original SIMS items that could accurately distinguish between consistent participants, symptom accentuators, and symptom producers in real personal injury cases. Subjects were personal injury claimants who had undergone forensic assessment, which is known to incentivize malingering and symptom accentuation. Item selection yielded short versions of the scale with as few as 8 items (to differentiate between consistent participants and symptom producers) and as many as 10 items (to differentiate between consistent and inconsistent participants). The scales had higher classification accuracy than the original SIMS and did not show the bias that was originally reported between false positives and false negatives.

Keywords SIMS · Psychic damage · Malingering · Machine learning · Feature selection

Introduction

Malingering is the dishonest and intentional production or exaggeration of physical or psychological symptoms in order to obtain external gain (Tracy & Rix, 2017). Although malingering is coded in both the ICD-11 (World Health Organization, 2019) and the DSM-5 (American Psychiatric Association, 2013), it is not a binary "present" or "absent" phenomenon: it may exist in specific domains (e.g., psychological, cognitive, and medical domains), it is often comorbid with formal disorders (Mazza et al., 2019c; Rogers & Bender, 2018), and it can be classified into several types (Akca et al., 2020; Lipman, 1962; Resnick, 1997). Due to the considerable variation produced by these nuances, it is difficult to measure the prevalence of malingering in clinical and forensic populations. According to forensic practitioners, malingering likely occurs in 15–17% of forensic cases (Rogers & Bender, 2018; Young, 2014). However, some studies have estimated a much higher prevalence, especially in forensic and non-forensic neuropsychological settings, with approximate rates ranging from 30 to 50% (Ardolf et al., 2007; Chafetz, 2008; Larrabee et al., 2009; Martin & Schroeder, 2020; Mittenberg et al., 2002). Given the cost and implications of malingering to the healthcare system, it is not surprising that several instruments have been introduced to assess the credibility of symptom presentations (see, e.g., the Structured Interview of Reported Symptoms—Second edition [SIRS-2; Rogers et al., 2010], the Test of Memory Malingering [TOMM; Tombaugh, 1996], the Self-Report Symptom Inventory [SRSI; Merten et al., 2016], and the Inventory of Problems-29 [IOP-29; Viglione et al., 2017; see also Roma et al., 2019a]). Among these instruments, the most widely used standalone symptom validity test in Europe and North America is the Structured Inventory of Malingered Symptomatology (SIMS; Dandachi-FitzGerald et al., 2013; Martin et al., 2015; Smith & Burger, 1997).

✉ Cristina Mazza
mazzacristina87@gmail.com

[1] Department of Surgical, Medical Molecular & Critical Area Pathology, University of Pisa, Pisa, Italy

[2] Department of Neuroscience, Imaging and Clinical Sciences, G. D'Annunzio University, Chieti-Pescara, Italy

[3] Department of General Psychology, University of Padova, Padova, Italy

[4] Department of Human Neuroscience, Sapienza University of Rome, Rome, Italy

Published online: 13 October 2020

The SIMS is a multi-axial self-report questionnaire that has been validated with clinical-forensic, psychiatric, and non-clinical populations. It is composed of a list of 75 implausible symptoms or statements that subjects must endorse or reject. It relies on the principle that malingerers may not know which symptoms truly characterize a given psychopathological condition, and they are thus likely to declare themselves as presenting with many atypical and rare psychopathological features. The SIMS covers a broad spectrum of pseudo-psychopathology. Its items index atypical depression, improbable memory problems, unlikely pseudo-neurological symptoms, doubtful claims of psychotic experiences, and hyperbolic signs of mental retardation. Each of these five categories (relating to neurologic impairment, affective disorders, psychosis, low intelligence, and amnestic disorders) is represented by a subscale composed of 15 items. The total number of implausible symptoms endorsed by a respondent represents the SIMS Total Score, which is the main symptom validity scale. The authors of the measure warned that SIMS subscales are not suitable for detecting feigned psychopathology and they only serve to evaluate which type of psychopathology a respondent is trying to feign, once it has been established that the Total Score exceeds the cutoff. A recent meta-analysis of 10 known groups and 24 simulation studies conducted by van Impelen et al., (2014) supported the efficacy and utility of the SIMS in forensic and clinical settings, despite some concerns with regard to its specificity when the commonly employed cutoff scores (i.e., ≥ 15 and ≥ 17) are used. The SIMS has been demonstrated to be fairly effective in discriminating between feigning and honest respondents, with effect sizes (Cohen's d) ranging from 1.1 to 3.0. The scale's sensitivity for the commonly employed cutoff scores has also been found to be adequate, ranging from 0.75 to 1.00, with corresponding specificity rates that are highly divergent (range 0.37–0.93), yet often alarmingly low.

Although the SIMS is frequently used in clinical and forensic assessments, its length—and therefore the time required for its administration—can be problematic. For this reason, it would be beneficial for clinicians to have brief measures to detect possible exaggeration or symptom production (Edens et al., 2007). With the aim of developing a brief and reduced item version of the SIMS, Malcore et al., (2015) performed a comprehensive item analysis that produced an abbreviated version of the SIMS, composed of 37 items and four (vs. five) subscales (Neurologic Impairment, Affective Disorders, Psychosis, and Amnestic Disorders), which maintained the integrity of the original SIMS. With the same intent of Malcore et al., (2015), we sought to investigate whether a reduced and easier to administer version of the SIMS could be developed using new machine learning (ML) techniques.

It has been suggested that the performance of ML algorithms may compare favorably with that of standard psychometric techniques when it comes to item analysis and test construction (Mazza et al., 2019b; Orrù et al., 2020a,Orrù et al., 2020b; Pace et al., 2019). Machine learning (ML) algorithms are usually trained and validated on an initial sample of data to make predictions on a completely new set of data (the test set) without being explicitly programmed to do so. The technique has been used to distinguish between feigners and honest respondents in a variety of settings. For instance, ML has shown extremely promising accuracy with regard to the detection of false identities (Monaro et al., 2018a, b), feigned depression (Monaro et al., 2018a, b), and feigned amnesia (Zago et al., 2019), and it has even been shown to be successful in detecting intentional underrepresentation of psychopathology (i.e., *faking good*; Burla et al., 2019; Mazza et al., 2019a, 2020, 2019b; Roma et al., 2014, 2018, 2019b, 2019c; Roma et al., 2016).

Recently, Mazza et al. (2019c) employed ML in an attempt to identify a strategy to distinguish between symptom accentuators, symptom producers, and consistent (i.e., truth telling) participants. Neglecting the diversity of malingering expressions, the researchers mostly considered it a unitary construct, both theoretically and empirically, assimilating aspects of both symptom production and symptom accentuation. In more detail, they analyzed the SIMS (Smith & Burger, 1997) and the Minnesota Multiphasic Personality Inventory-2 Restructured Form (MMPI-2-RF; Ben-Porath & Tellegen, 2008) validity scales of 132 subjects with a diagnosed adjustment disorder with mixed anxiety and depressed mood, who had undergone an assessment for psychiatric/psychological damage. It must be said that, in lie deception research (McCarthy-Jones and Resnick, 2014; Musso & Gouvier, 2014; Tracy & Rix, 2017), adjustment disorders have received less attention than other mental disorders (i.e., post-traumatic stress disorder, adult attention-deficit hyperactivity disorder, somatoform and dissociative disorders, psychosis), even though, in medico-legal contexts, disorders associated with depression and anxiety (e.g., chronic adjustment disorder with mixed anxiety and depressed mood) are the most frequently simulated (Mittenberg et al., 2002), at a rate of over 50% (Santamaría et al., 2013). The results indicated that the SIMS Total Score, scores for the Neurologic Impairment and Low Intelligence subscales, and scores for the MMPI-2-RF Infrequent Responses and Response Bias subscales successfully discriminated between symptom accentuators, symptom producers, and consistent participants. ML was used to identify the most effective parameter for classifying the three groups, recognizing SIMS Total Score as the best indicator.

In the present study, we extended the results reported by Mazza et al. (2019c) by investigating whether ML techniques could be used to develop a valid and reliable SIMS short

Table 1 Group composition according to the criteria

	Consistent participants	Symptom accentuators	Symptom producers
Criterion 1	Congruent	Congruent	Congruent or incongruent
Criterion 2	Congruent	Congruent or incongruent	Congruent or incongruent
Criterion 3	Congruent	Congruent or incongruent	Congruent or incongruent

form. Classic psychometrics and Rasch models (including item response theory [IRT]) (Bond, & Fox, 2015) treat selected items as local estimators of individual features. In classic psychometrics, linear correlation is usually the base for estimating item relevance in group discrimination tasks. In IRT, an item may be used to rank both subject ability and item difficulty. In classical test theory, a common strategy for abbreviating inventories is to select a subset of items from each scale that maximizes correlations between the item and the total score, while maintaining high internal consistency (e.g., Cox & Alexander, 1995; Goldberg et al., 2006; Lang & Stein,2005; Troldahl & Powell, 1965). Similarly, IRT is used to select a subset of items for a shortened scale that replicates the performance of the full test (e.g., Embretson & Reise, 2000), including items that are sensitive to the full score range. Techniques such as these, which assess the value of individual items, may miss the combined boosting effect of features that, considered in isolation, may not seem important for group discrimination.

In the present research, we sought to develop shorter versions of the SIMS using ML feature selection models. It has been shown that reducing the number of predictors in ML classifiers may increase accuracy (by approximately 15%) by eliminating redundant predictors (Karabulut et al., 2012). Our approach primarily focused on prediction accuracy (Orrù et al., 2020a; Yarkoni & Westfall, 2017), maximum generalizability, and cross-validated results, rather than statistical model fit to the data (which is the main focus of IRT and classical approaches). Our rationale for this was that the classical focus on model fit frequently overfits, with the consequence that non–cross-validated results still replicate.

Method

Participants and Procedure

The SIMS Total Scores of 132 participants, collected by Mazza et al. (2019c) in their recently published study, were re-analyzed to develop the short version of the SIMS reported here. Specifically, Mazza et al. (2019c) collected data on 132 participants who had undergone—between January and December 2018—a court ordered mental health examination in the context of a lawsuit involving psychological damage. All were diagnosed with an adjustment disorder with mixed anxiety and depressed mood (309.28). According to a specific three-phase procedure, participants were divided into three groups: consistent participants ($N = 49$), symptom accentuators ($N = 44$), and symptom producers ($N = 39$). Participants were considered consistent if they had (1) submitted suitable documentation (e.g., an illness certificate for work) that was judged congruent/coherent with their diagnosis of an adjustment disorder with anxiety and depressed mood, (2) showed symptoms compatible with the aforementioned diagnosis, and (3) referred to impaired psychological and psychosocial functioning as an effect of the adjustment disorder with anxiety and depressed mood. Participants were labelled symptom accentuators if they were judged incongruent/incoherent with respect to *either* criterion 2 or 3 and, consequently, had shown an inflated manifestation of clinical and emotional symptoms or an inflated impairment in day-to-day functioning (in social, working, and other important areas) due to their adjustment disorder with anxiety and depressed mood. Finally, participants were considered symptom producers if they were judged incongruent/incoherent with respect to *at least* two of the criteria described above (see Table 1).

The three groups differed in age ($F(2, 129) = 8.373$, $p < 0.001$) and educational level ($F(2, 129) = 4.240$, $p = 0.016$), but not gender (χ^2 (2) = 3.341, $p = 0.188$). For more detail on the participants and procedure, please refer to Mazza et al. (2019c).

Data Analyses: Cross-validation

A tenfold cross-validation procedure was used for all reported analyses. Cross-validation is usually very effective at measuring the exact replication of a result (Cumming, 2008). Exact replication refers to replication in which all conditions of the original experiment are maintained, but it does not address the replicability of the main finding following the introduction of minor variations. As cross-validation consists of evaluating models on a hold-out set of experimental examples, the set does not differ from the examples used for model development. Cross-validation estimates true performance while preventing model overfit. For this reason, it is a compulsory step in ML analysis, though its use in the analysis of psychological experiments is limited. There are a number of cross-validation procedures, but stratified tenfold cross-validation has been found to be especially effective at approximating out-of-sample results (James et al., 2013). In order to develop models that are

able to generalize new (unseen) data, the procedure should (1) remove 20% of the data for validation; (2) run tenfold cross-validation on the remaining 80% of the data, with the aim of selecting optimal parameters; (3) train the model with all 80% of the data with optimal parameters; and (4) test the model on the 20% validation set. The result of step 4 should provide the best approximation of exact replication. In the present study, all reported cross-validation and ML analyses were run using WEKA (Frank et al., 2016).

ML Feature (Item) Selection

For the purpose of this research, feature selection can be considered synonymous with item selection. Feature selection is the process by which the best (i.e., most accurate) subset of items is automatically selected. In the present case, accuracy referred to accurate discrimination between consistent participants, symptom accentuators, and symptom producers in a personal injury setting. In feature selection, the search space of item subsets is discrete and consists of all possible combinations of items. The objective is to identify the best—or a good enough—combination of items that improves on or demonstrates equal performance to that of the complete and original scale. Two key benefits of ML feature selection include reduced overfitting (as less redundant data reduces the risk of decisions being based on noise) and improved accuracy (as less misleading data improves model accuracy) (Karabulut et al., 2012).

Results

First, we report on the classification of participants in the three groups (consistent participants, symptom accentuators, and symptom producers). Second, we report on the classification of participants in the two extreme groups. Third—and finally—we report on the classification of participants into groups of consistent versus inconsistent participants (with the latter composed of all symptom accentuators and symptom producers).

Classification in Three Groups

The dataset comprised three groups: consistent participants ($N=49$), symptom accentuators ($N=44$), and symptom producers ($N=39$). A preliminary multi-class classifier[1] was developed using Naïve Bayes. Given that the size of the three groups was unequal, it was important to establish a base classification accuracy for the purpose of comparison.

Table 2 Confusion matrix of the Naïve Bayes multi-class classifier. Most classification errors pertained to symptom accentuators, with individuals misclassified as consistent participants or symptom producers

Actual classified as	Consistent participants	Symptom accentuators	Symptom producers
Consistent participants	43	17	4
Symptom accentuators	3	11	10
Symptom producers	3	16	25
Total	49	44	39

For this, we used ZeroR, which is the simplest classification method; it relies on the target and ignores all predictors. The ZeroR classifier predicts the majority category (class). Although ZeroR has no predictability power, it is useful for determining baseline performance. In the present analysis, the result of this calculation was 37.12%; this was regarded as the benchmark efficiency for all of the multi-class classifiers reported below. When classifying into three groups, Naïve Bayes yielded 59.8% classification accuracy with an AUC = 0.704 (see Table 2).

The OneR algorithm[2] generated comparable classification accuracy (60.6%; AUC = 0.703). The best performing rule boundaries (identified using tenfold cross-validation) were as follows: $IF\ SIMS < 14 = consistent\ participants$, $IF\ 14 < SIMS < 19 = symptom\ accentuators$, and $IF\ SIMS > 19 = symptom\ producers$. The OneR algorithm corresponds to the hand picking by visual inspection cutoff identified in previous research on the SIMS. It is clear from the confusion matrix reported in Table 2 that symptom accentuators were hardest to correctly classify, as they were frequently misclassified as consistent participants or symptom producers. This result indicates that the original expert-based classification captured the complexity of distinguishing symptom accentuators from consistent participants, on the one hand, and symptom accentuators from symptom producers, on the other.

Consistent Participants Versus Symptom Producers

For the reasons reported above, in order to isolate the items that more efficiently distinguished between consistent participants and symptom producers, we excluded symptom accentuators from the analyses reported below. As a starting point, we evaluated the full 75-item SIMS in order to determine a benchmark for comparing the results of the shortened version of the scale. In our dataset, the overall classification accuracy using the suggested cutoff of 14 (van Impelen

[1] A multi-class classifier classifies instances into more than two classes.

[2] The OneR algorithm identifies the best performing single decision rule.

et al., 2014) yielded an overall accuracy of 87.5% (correct classification in 77 out of 88 instances), with six consistent participants scoring ≥ 15 and five symptom producers scoring < 15. This preliminary evaluation indicated that the full 75-item SIMS, when applied to our dataset, resulted in a maximum accuracy of approximately 90%, with no significant difference between sensitivity and specificity.

Any new and more efficient shorter version of the SIMS would have to demonstrate performance of at least 87.5%. As regards ML item reduction, it is well known (Chu et al., 2012; Karabulut et al., 2012) that ML classification accuracy benefits from the elimination of redundant features (in the present case, test items). For this reason, we expected that a short version of the SIMS would not underperform relative to the full version.

Item Selection Strategies

In order to determine whether subsets of items could demonstrate performance at least as high as the standard 75-item SIMS scale when distinguishing between consistent participants and symptom producers, we used two different feature selection strategies: filter and wrapper. Filter methods evaluate the individual contribution of attributes. Wrapper methods, in contrast, systematically analyze all combinations of features in order to isolate the particular combination that maximizes discrimination between classes. Wrapper methods are linked to a base classifier, and they measure the "usefulness" of features based on classifier performance. Filter methods, however, are based on features' intrinsic properties (i.e., "relevance"), which are measured via univariate statistics rather than cross-validation.

In the present research, we developed two shortened versions of the SIMS using filter and wrapper feature selection procedures, respectively. We then checked whether these shortened versions demonstrated at least equal performance to the original and full 75-item SIMS in terms of the ability to accurately distinguish between consistent participants and symptom producers.

Correlation-Based Filter Method

The correlation-based filter method selected the eight most highly correlated items with each group (consistent participants vs. symptom producers) with *r* values ranging from 0.424 to 0.508.

As shown in Table 3, these highly correlated items, when used as ML classifier predictors, achieved at least the same classification accuracy as the full SIMS scale. In order to compare the ML selected items with those selected by Malcore et al., (2015) using more traditional methods of item selection, we also calculated the results on the same data and same classifiers, using the 37-item score.

Table 3 Performance of six ML models using the full 75-item scale as input or the shortened scale composed of the eight most highly correlated items with the two classes (consistent participants vs. symptom producers). Irrespective of the classifier, all 8-item shortened scales performed better than or equal to the original 75-item SIMS. The table also reports the accuracies obtained, replicating the exact conditions used for the Malcore et al. (2015) 37-item short version of the SIMS

Classifier	SIMS 75-item	SIMS 8-item	SIMS 37-item (Malcore et al. 2015)
Naïve Bayes	82%	92%	83%
Logistics	73%	94%	77%
SVM	84%	94%	78%
Random forest	88%	88%	86%
PART	89%	90%	81%

Rather than predicting class values, it is sometimes convenient to predict the probability of an observation belonging to each possible class. A classifier is well calibrated when the predicted probabilities correspond to the observed probabilities. We calculated the calibration curve for the 75- and 8-item SIMS. Not only was the 8-item SIMS more accurate, but it was also well calibrated, as demonstrated by the visual comparison of the calibration curves for the Naïve Bayes classifier (Figs. 1 and 2). Similar results were also observed for the other classifiers.

ML models are usually regarded as opaque solutions, as their decision logic can be unclear. Thus, recent research has focused on developing explainable and intuitively transparent models (Lundberg et al., 2020). Some ML models can be fine-tuned to more easily convey the decision rules. In the present study, in order to derive an easy-to-understand set of decision rules, we analyzed the performance of the PART algorithm (Frank & Witten, 1998). Incidentally, it should be noted that the complex decision rule derived by the PART algorithm did not use all of the original eight items but succeeded in maximizing classification accuracy using only six of the original eight items. In short, it performed item selection within a preliminary item selection. This discrepancy was the cost of having an interpretable model.

Effect of Class Imbalance

In the present research, class imbalance was reflected in the fact that there were 49 consistent participants but only 39 symptom producers. Given that class imbalance is regarded as a source of reduced classification accuracy in ML, we sought to evaluate its impact (Hasanin et al., 2019) in our analyses. In order to balance classes, we undersampled the majority class by randomly extracting 10 cases from the group of consistent participants in order to match the group's size to that of the symptom producers. Re-running all of

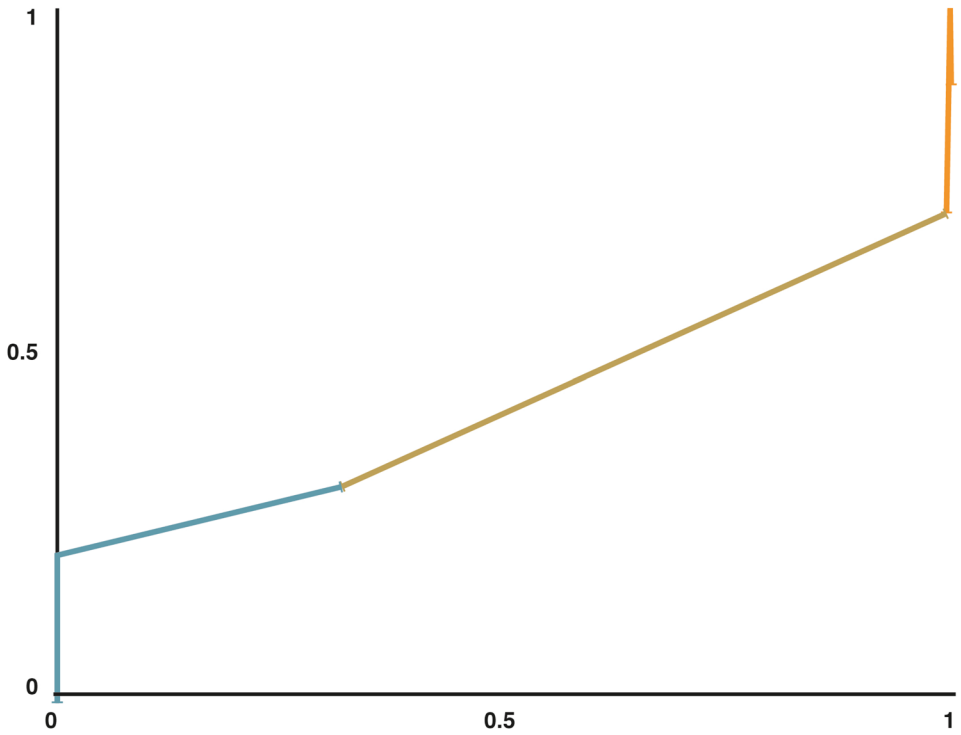

Fig. 1 Calibration curve for the 75-item SIMS. Overall accuracy of the Naïve Bayes classifier was 82%. The perfectly calibrated classifier corresponds to the diagonal

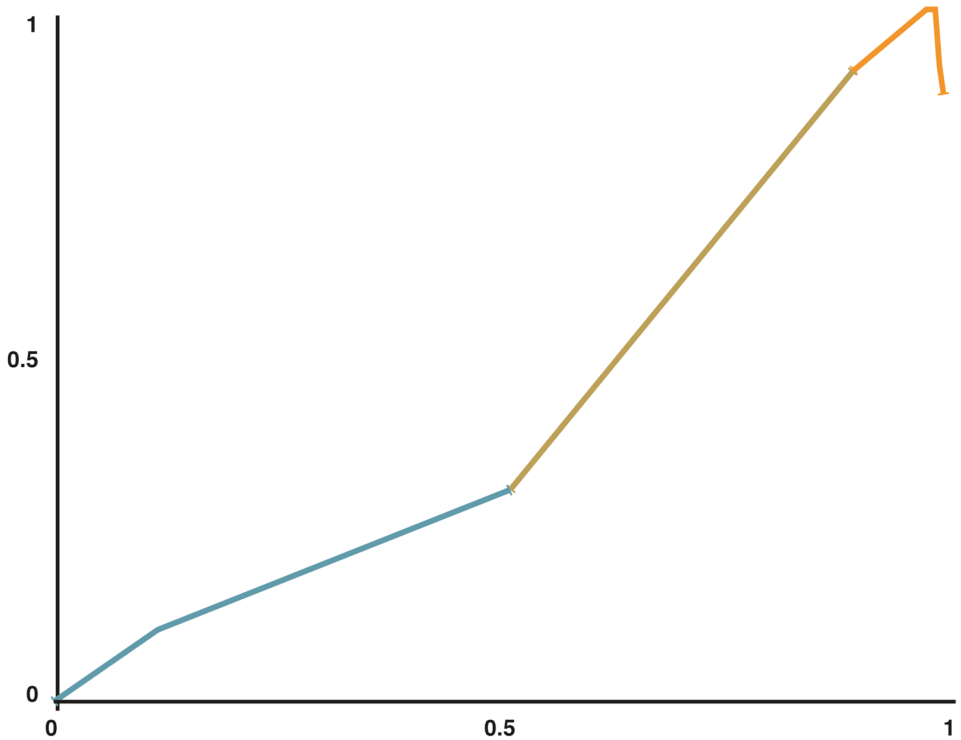

Fig. 2 Calibration curve for the 8-item SIMS. Overall accuracy of the Naïve Bayes classifier was 92%. The calibration curve of this short version clearly overlaps the diagonal better than the original SIMS test

the classifiers using the subset of items derived through the correlation analysis generated results that overlapped those reported with the class imbalance (Table 4).

This result indicates that, in our specific case, the class imbalance did not reduce classification accuracy. For this reason, all analyses reported below were run on the original imbalanced dataset of 49 consistent participants and 39 symptom producers.

Table 4 Performance of the five ML models using the aforementioned eight items as input with imbalanced and balanced data

Classifiers	Imbalanced data	Balanced data
Naïve Bayes	92%	94%
Logistics	94%	95%
SVM	94%	94%
Random forest	88%	91%

Table 5 Performance of six ML models using the full 75-item scale as input or the eight items selected using a wrapper with Naïve Bayes. All classifiers generated good performance with the reduced item set

Classifier	Input = 75 items	Input = 8 items
Naïve Bayes	82%	94%
Logistics	73%	93%
SVM	84%	94%
Random forest	88%	90%
J48	82%	90%
PART	89%	87%

Feature Selection with Wrapper Methods

While wrapper methods essentially solve the "real" problem (optimizing classifier performance), they are computationally more expensive than filter methods due to the repeated learning steps and cross-validation. However, in most cases, they yield better results than filter methods (Kohavi & John, 1997). As already mentioned, the optimal features that maximize the results when wrapper methods are used depend on the classifier one wishes to optimize. In order to develop features optimized for an interpretable classifier, we used Naïve Bayes as a base learner. When the wrapper was applied, 8 items from the original SIMS were selected. Overall accuracy was 94.3% (with correct classification of 48 out of 49 consistent participants and 35 out of 39 symptom producers). The correlation of the Total Score of the 8-item SIMS with the resulting category was 0.731, while the corresponding figure for the full SIMS Total Score was 0.769. We also verified that performance of the selected items did not reduce when other classifiers were used (Table 5).

As already noted, most of the classifiers reported in Table 6 have a function that is difficult to understand. However, some of the models, as decision rules, may be intuitively understood; for this reason, they are better suited to direct application in clinical practice (Arrieta et al., 2020). The PART classifier is one such decision rule algorithm and, in this case, using the eight features reported above as input, it yielded an overall accuracy of 89%. The specific application of this algorithm found two items among the original

Table 6 Correlations matrix between the group classification, the SIMS under investigation, and the selected MMPI-2 validity scale

		Groups	SIMS_8_FILTER	SIMS_8_WRAPPER	SIMS_10_WRAPPER	SIMS_75_STANDARD	SIMS_MALCORE et al. (2015)	F-r	Fp-r	Fs	FBS-r	RBS
Groups	r_{pb}	-	0.618**	0.587**	0.611**	0.603**	0.524**	0.469**	0.320**	0.368**	0.381**	0.446**
SIMS_8_FILTER	*r*	0.618**	-	0.736**	0.778**	0.790**	0.722**	0.402**	0.209*	0.309**	0.305**	0.366**
SIMS_8_WRAPPER	*r*	0.587**	0.736**	-	0.692**	0.778**	0.690**	0.344**	0.214*	0.297**	0.231**	0.305**
SIMS_10_WRAPPER	*r*	0.611**	0.778**	0.692**	-	0.808**	0.766**	0.433**	0.301**	0.294**	0.306**	0.381**
SIMS_75_STANDARD	*r*	0.603**	0.790**	0.778**	0.808**	-	0.930**	0.462**	0.234**	0.377**	0.381**	0.436**
SIMS_MALCORE et al. (2015)	*r*	0.524**	0.722**	0.690**	0.766**	0.930**	-	0.460**	0.229**	0.353**	0.367**	0.403**
F-r	*r*	0.469**	0.402**	0.344**	0.433**	0.462**	0.460**	-	0.623**	0.652**	0.513**	0.671**
Fp-r	*r*	0.320**	0.209*	0.214*	0.301**	0.234**	0.229**	0.623**	-	0.520**	0.296**	0.474**
Fs	*r*	0.368**	0.309**	0.297**	0.294**	0.377**	0.353**	0.652**	0.520**	-	0.695**	0.755**
FBS-r	*r*	0.381**	0.305**	0.231**	0.306**	0.381**	0.367**	0.513**	0.296**	0.695**	-	0.772**
RBS	*r*	0.446**	0.366**	0.305**	0.381**	0.436**	0.403**	0.671**	0.474**	0.755**	0.772**	-

"Groups" is classified as a dummy variable (0 = consistent participants, 1 = inconsistent participants)

*$p < 0.05$; **$p < 0.01$

eight that were redundant and therefore not included in the rule.

Of note, the wrapper method selected a subset of high-performance attributes that, taken in isolation, did not all demonstrate maximum correlation. As already mentioned, the wrapper technique explores all possible combinations of features (all pairs, triplets, etc.) and selects the optimal subset that, in tenfold cross-validation, results in the best performance. In the present case, the individual items selected by the wrapper using Naïve Bayes as a base learner reported correlations ranging from 0.01 to 0.47. Only 3 items were among the 10 with the highest correlations. Further, it is relevant to note that an item was selected even though its correlation was virtually 0; nonetheless, when combined with other items, it contributed to the maximum discrimination between consistent participants and symptom producers.

Items Distinguishing Between Consistent and Inconsistent Participants (i.e., Symptom Accentuators and Symptom Producers)

The above analyses referred to an eight-item version of the SIMS that maximized the discrimination between consistent participants and symptom producers (with symptom accentuators excluded from the analysis). We also conducted similar analyses contrasting consistent participants with inconsistent participants (composed of all symptom accentuators and symptom producers). Given that the size of the two groups significantly differed (consistent participants = 49; inconsistent participants = 83), we first balanced the two classes by undersampling the majority class. In this case, the inconsistent group had the higher number of participants, so we randomly reduced it to achieve the same size as the consistent group.

Similar to the previous analyses, we selected the best features using a wrapper method with Naïve Bayes as a base classifier. The results indicated 10 items, with correlations *r* ranging from 0.13 to 0.43 The Naïve Bayes classifier using 10 items yielded an overall classification accuracy of 91.8%. Only 5 out of the 45 consistent participants were wrongly classified, and a similar result was observed for the inconsistent participants. It is interesting to note that, also in this case, the selected features were not exclusively those with the highest correlations with the output (consistent vs. inconsistent participants). Rather, only 5 of the selected features were among the 10 with the highest correlations. Note that the PART algorithm does not select 2 of the 10 originally selected items. This yields a reduction in accuracy (87%) with respect to interpretable algorithms such as SVM (94%). This reduction in accuracy for interpretable models is usually observed relative to higher accuracies for uninterpretable models and is regarded the cost of interpretability (Lundberg et al., 2020).

Considering copyright restrictions, the newly introduced SIMS-8 and SIMS-10 are available to researchers via e-mail only, upon reasonable request.

Convergent Validity

Finally, to test the convergent validity of the resulting three brief forms of the SIMS, we examined their correlations with the SIMS standard version (i.e., SIMS 75-item), the version proposed by Malcore et al., (2015), and the MMPI-2-RF (Ben-Porath & Tellegen, 2008) subscales designed to detect overreporting and response bias (i.e., infrequent responses, F-r; infrequent psychopathology responses, Fp-r; symptom validity, FBS-r; infrequent somatic responses, Fs; response bias, RBS). We also tested the point-biserial correlations between scores on the aforementioned scales and participants' classification as consistent vs. inconsistent. All correlations were positive and significant, with coefficients ranging from weak to strong, which is indicative of adequate convergent validity (Table 6). In more detail, the results revealed moderate positive correlations between group classification and the F-r and RBS scales, which a recent meta-analysis (Sharf et al., 2017) indicated as the most sensitive scales for feigned mental disorders (RBS 0.93, cutoff ≥ 80; F-r 0.71, cutoff ≥ 10), together with the FBS-r (0.84, cutoff ≥ 80).

Discussion

In the present research, we applied ML techniques to derive SIMS subtests that, with as few as 8–10 items, could achieve a classification accuracy similar to that obtained by the full 75-item scale when differentiating between consistent and inconsistent participants. The "standard" SIMS consists of 75 true–false bizarre items that span five symptom domains (i.e., psychosis, neurology, amnesia, mental disability, affective disorders). While it has reasonable psychometric properties (e.g., high sensitivity), it is also subject to a number of limitations (see, e.g., van Impelen et al., 2014), including the length. In the present study, we used the SIMS Total Scores of 132 participants, collected by Mazza et al., (2019c) in their recently published study, to develop shorter and easier to use versions of the scale. Participants had undergone a court ordered mental health examination in the context of a lawsuit involving psychological damage; therefore, their propensity to malinger was triggered by a real compensation setting. All participants had been diagnosed with an adjustment disorder with mixed anxiety and depressed mood (309.28) (APA, 2013). According to a specific three-phase procedure detailed in the original article, they were divided into three groups: consistent participants ($N = 49$), symptom accentuators ($N = 44$), and symptom producers ($N = 39$). In

our disability claiming forensic sample, the suggested SIMS Total Score cutoff ≥ 15 yielded an accuracy of 87.5% in sorting between consistent participants and symptom producers; this figure was in line with previous results (van Impelen et al., 2014).

In selecting the most informative items of the SIMS, we used ML rather than standard psychometric methods of item analysis (e.g., classical test theory or IRT). ML treats the data as unknown and mainly focuses on classification accuracy and the generalization of results, rather than statistical model fit (as in IRT). In this way, ML prediction, achieved using general purpose learning algorithms to find patterns in often numerous and highly complex datasets, aims at forecasting unobserved outcomes or future behavior. In short, ML models are "model agnostic" and focused on prediction (Orrù et al., 2020a); generally speaking, tests in clinical and forensic practice share this goal.

The objective of the present investigation was to reduce the number of items while maintaining classification accuracy. ML feature reduction reduces the number of predictors (in our case, SIMS items); within this process, it is frequently observed that the elimination of redundant features (i.e., items) increases classification accuracy (Kohavi & John, 1997). The feature selection methodology employed here distinguished between filter and wrapper methods. Filter methods evaluate the value of individual features and retain only the best features. One such method uses correlation values, as in classical item analysis. In contrast, wrapper methods check all possible subsets of items (i.e., pairs, triplets, etc.) and identify the most effective of these subsets. These methods often show that the best performing item subset is not necessarily composed of all of the best performing individual items. This was also our observation in the model built using a wrapper technique.

It is frequently observed that, in most datasets analyzed using ML models, similar prediction accuracies are achieved using models that rely on very different assumptions. This was also verified in the present study (e.g., the support vector machine, Naïve Bayes, and random forest methods resulted in similar accuracies). Well performing models are often difficult to interpret, giving rise to a clear interpretability/accuracy trade-off. For example, Fernandez-Delgado et al. (2014) evaluated the performance of 179 ML classifiers on 121 datasets, concluding that random forest and support vector machine (SVM) (Orrù et al., 2012) classifiers achieved the top performance (with no significant difference between the two). Random forest, neural network, and SVM classifiers are all difficult to interpret. Simpler models, such as pruned decision rules and Naïve Bayes models, are easier to interpret but rarely result in the best performance. In order to increase the interpretability of our results, we also reported ML models based on PART-derived decision rules. The classification accuracy of these decision rules was slightly lower but had the benefit of greater interpretability and could be better comprehended by clinicians.

A further positive result of the ML item analysis was that no specific difference was found between false positives and false negatives. By contrast, a meta-analysis of the 75-item SIMS indicated that the scale's specificity (i.e., correct identification of consistent participants) may be alarmingly low (van Impelen et al., 2014). All of the short versions of the scale presented here showed no differences in their rates of false negatives versus false positives. Furthermore, all of the short versions we identified had comparable levels of accuracy for both sensitivity and specificity and performed as well as the full scale in discriminating consistent from inconsistent participants. The convergent validity between the three SIMS versions investigated here and the selected MMPI-2-RF validity scales also supported the use of the brief versions.

The reader should bear in mind that no expert evaluates the credibility of respondents' symptoms on the basis of a single measure, alone—whether that measure is the full SIMS or a shortened version of the scale. Rather, these instruments are designed to merely support clinical decisions with evidence-based results. Because the 75-item SIMS requires extended administration time, brief and easy-to-use measures to detect the feigning of mental disorders could be usefully introduced to clinical settings, as they may yield similar information to the full scale.

Strengths and Limitations

The aim of the present study was to overcome one of the main limitations of the SIMS by reducing its number of items through the use of ML models and, consequently, reducing the duration of its administration. The three short versions of the SIMS reported here were built using a bottom-up approach, and this is both a strength and a weakness of our study. In fact, our results are based on participants recruited in a real forensic setting who were not instructed to feign a mental disorder within an experimental paradigm. Thus, subjects' classification as consistent responders, symptom accentuators, and symptom producers was based on the assessment of clinicians and not on objective measures or a notion of a "gold standard." Furthermore, while our complex SIMS classification rules may result in making the test difficult to feign without detection, they require specific computer software to administer.

One important future direction of our work will be to compare the classification accuracy of the three brief and easy-to-use SIMS versions reported here with other criterion variables (i.e., IOP-29), both in experimental and in forensic settings, in order to also identify symptom accentuators, who comprise the majority in clinical and forensic practice.

Authors' Contributions Conceived the experiment: GO and GS. Data acquisition: PR. Data analysis: GO and SF. Data interpretation: all authors. Drafting of the manuscript: GO and CM. All authors revised the manuscript critically and gave final approval of the version to be published.

Funding Open access funding provided by Università degli Studi G. D'Annunzio Chieti Pescara within the CRUI-CARE Agreement.. Open access funding provided by Università degli Studi G. D'Annunzio Chieti Pescara within the CRUI-CARE Agreement.

Data Availability The data that support the findings of this study are available from the corresponding author (C.M.), upon reasonable request.

Declarations

Conflict of Interest The authors declare that they have no conflicts of interest.

Ethics Approval This study was carried out with written informed consent by all subjects, in accordance with the Declaration of Helsinki. It was approved by the local ethics committee (Board of the Department of Human Neuroscience, Faculty of Medicine and Dentistry, Sapienza University of Rome).

References

Akca, Y. A., Slootmaekers, L., & Boskovic, I. (2020) Verifiability and symptom endorsement in genuine, exaggerated, and malingered pain. *Psychological Injury and Law,* 1–11. 13 p.235-245 https://doi.org/10.1007/s12207-02009375-w

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorder* (5th ed.). Washington, DC: Author.

Ardolf, B. R., Denney, R. L., & Houston, C. M. (2007). Base rates of negative response bias and malingered neurocognitive dysfunction among criminal defendants referred for neuropsychological evaluation. *The Clinical Neuropsychologist, 21*(6), 899–916. https://doi.org/10.1080/13825580600966391

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., & Chatila, R. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58,* 82–115.

Ben-Porath, Y. S., & Tellegen, A. (2008). *MMPI-2-RF: Manual for administration, scoring and interpretation*. Minneapolis, MN: University of Minnesota Press.

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Abingdon-on-Thames, England, UK: Routledge.

Burla, F., Mazza, C., Cosmo, C., Barchielli, B., Marchetti, D., Verrocchio, M.C., & Roma, P. (2019). Use of the Parents Preference Test in child custody evaluations: Preliminary development of Conforming Parenting Index. *Mediterranean Journal of Clinical Psychology*, *7*(3). https://doi.org/10.6092/2282-1619/2019.7.2213.

Chafetz, M. D. (2008). Malingering on the social security disability consultative exam: Predictors and base rates. *The Clinical Neuropsychologist, 22*(3), 529–546. https://doi.org/10.1080/13854040701346104

Chu, C., Hsu, A. L., Chou, K. H., Bandettini, P., & Lin, C. P. (2012). Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *Neuroimage, 60*(1), 59–70. https://doi.org/10.1016/j.neuroimage.2011.11.066

Cox, R. M., & Alexander, G. C. (1995). The abbreviated profile of hearing aid benefit. *Ear and Hearing*, *16*(2), 176–186. Retrieved from https://journals.lww.com/ear-hearing/pages/default.aspx

Cumming, G. (2008). Replication and p intervals: P values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3*(4), 286–300. https://doi.org/10.1111/j.1745-6924.2008.00079.x

Dandachi-FitzGerald, B., Ponds, R. W., & Merten, T. (2013). Symptom validity and neuropsychological assessment: A survey of practices and beliefs of neuropsychologists in six European countries. *Archives of Clinical Neuropsychology, 28*(8), 771–783. https://doi.org/10.1093/arclin/act073

Edens, J. F., Poythress, N. G., & Watkins-Clay, M. M. (2007). Detection of malingering in psychiatric unit and general population prison inmates: A comparison of the PAI, SIMS, and SIRS. *Journal of Personality Assessment, 88*(1), 33–42. https://doi.org/10.1080/00223890709336832

Embretson, S. E., & Reise, S. P. (2000). *Multivariate applications books series. Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research, 15*, 3133–3181. Retrieved from https://www.jmlr.org

Frank, E., Hall, M. A., & Witten, I. H. (2016). *The WEKA workbench. Online appendix for Data mining: Practical machine learning tools and techniques* (4th ed.). San Francisco, CA: Morgan Kaufmann.

Frank, E., & Witten, I. H. (1998). *Generating accurate rule sets without global optimization* (working paper 98/2). Hamilton: University of Waikato, Department of Computer Science.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*(1), 84–96. https://doi.org/10.1016/j.jrp.2005.08.007

Hasanin, T., Khoshgoftaar, T. M., Leevy, J. L., & Seliya, N. (2019). Examining characteristics of predictive models with imbalanced big data. *Journal of Big Data, 6,* 69. https://doi.org/10.1186/s40537-019-0231-2

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, pp. 3–7). New York: Springer.

Karabulut, E. M., Özel, S. A., & Ibrikci, T. (2012). A comparative study on the effect of feature selection on classification accuracy. *Procedia Technology, 1,* 323–327. https://doi.org/10.1016/j.protcy.2012.02.068

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*(1–2), 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X

Lang, A. J., & Stein, M. B. (2005). An abbreviated PTSD checklist for use as a screening instrument in primary care. *Behaviour*

Research and Therapy, 43(5), 585–594. https://doi.org/10.1016/j.brat.2004.04.005

Larrabee, G. J., Millis, S. R., & Meyers, J. E. (2009). 40 plus or minus 10, a new magical number: Reply to Russell. https://doi.org/10.1080/13854040902796735

Lipman, F. D. (1962). Malingering in personal injury cases. *Temple Law Quarterly, 35*(2), 141–162. Retrieved from https://www.templelawreview.org/

Lundberg, S. M., Erion, G., Chen, H., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2,* 56–67. https://doi.org/10.1038/s42256-019-0138-9

Malcore, S. A., Schutte, C., Van Dyke, S. A., & Axelrod, B. N. (2015). The development of a reduced-item Structured Inventory of Malingered Symptomatology (SIMS). *Psychological Injury and Law, 8*(2), 95–99. https://doi.org/10.1007/s12207-015-9214-6

Martin, P. K., & Schroeder, R. W. (2020). Base rates of invalid test performance across clinical non-forensic contexts and settings. *Archives of Clinical Neuropsychology*. https://doi.org/10.1093/arclin/acaa017

Martin, P. K., Schroeder, R. W., & Odland, A. P. (2015). Neuropsychologists' validity testing beliefs and practices: A survey on North American professionals. *The Clinical Neuropsychologist, 29*(6), 741–776. https://doi.org/10.1080/13854046.2015.1087597

Mazza, C., Burla, F., Verrocchio, M.C., Marchetti, D., Di Domenico, A., Ferracuti, S., & Roma, P. (2019). MMPI 2-RF profiles in child custody litigants. *Frontiers in Psychiatry, 10.* https://doi.org/10.3389/fpsyt.2019.00725

Mazza, C., Monaro, M., Orrù, G., Burla, F., Colasanti, M., Ferracuti, S., & Roma, P. (2019b). Introducing machine learning to detect personality faking-good in a male sample: A new model based on MMPI-2-RF scales and reaction times. *Frontiers in Psychiatry*, *10*. https://doi.org/10.3389/fpsyt.2019.00389

Mazza, C., Monaro, M., Burla, F., Colasanti, M., Orrù, G., Ferracuti, S., & Roma, P. (2020). Use of mouse-tracking software to detect faking-good behavior on personality questionnaires: An explorative study. *Scientific Report, 10*, 4835. https://doi.org/10.1038/s41598-020-61636-5

Mazza, C., Orrù, G., Burla, F., Monaro, M., Ferracuti, S., Colasanti, M., & Roma, P. (2019c). Indicators to distinguish symptom accentuators from symptom producers in individuals with a diagnosed adjustment disorder: A pilot study on inconsistency subtypes using SIMS and MMPI-2-RF. *PLoS One*, *14*(12). doi:https://doi.org/10.1371/journal.pone.0227113

McCarthy-Jones, S., & Resnick, P. J. (2014). Listening to voices: The use of phenomenology to differentiate malingered from genuine auditory verbal hallucinations. *International Journal of Law and Psychiatry, 37*(2), 183–189. https://doi.org/10.1016/j.ijlp.2013.11.004

Merten, T., Merckelbach, H., Giger, P., & Stevens, A. (2016). The Self-Report Symptom Inventory (SRSI): A new instrument for the assessment of distorted symptom endorsement. *Psychological Injury and Law, 9*(2), 102–111. https://doi.org/10.1007/s12207-016-9257-3

Mittenberg, W., Patton, C., Canyock, E. M., & Condit, D. C. (2002). Base rates of malingering and symptom exaggeration. *Journal of Clinical and Experimental Neuropsychology, 24*(8), 1094–1102. https://doi.org/10.1076/jcen.24.8.1094.8379

Monaro, M., Gamberini, L., Zecchinato, F., & Sartori, G. (2018). False identity detection using complex sentences. *Frontiers in Psychology, 9,* 283. https://doi.org/10.3389/fpsyg.2018.00283

Monaro, M., Toncini, A., Ferracuti, S., Tessari, G., Vaccaro, M. G., De Fazio, P., & Sartori, G. (2018). The detection of malingering: A new tool to identify made-up depression. *Frontiers in Psychiatry, 9,* 249. https://doi.org/10.3389/fpsyt.2018.00249

Musso, M. W., & Gouvier, W. D. (2014). "Why is this so hard?" A review of detection of malingered ADHD in college students. *Journal of Attention Disorders, 18*(3), 186–201. https://doi.org/10.1177/1087054712441970

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

Orrù, G., Gemignani, A., Ciacchini, R., Bazzichi, L., & Conversano, C. (2020). Machine learning increases diagnosticity in psychometric evaluation of alexithymia in fibromyalgia. *Frontiers in Medicine, 6,* 319. https://doi.org/10.3389/fmed.2019.00319

Orrù, G., Monaro, M., Conversano, C., Gemignani, A., & Sartori, G. (2020). Machine learning in psychometrics and psychological research. *Frontiers in Psychology, 10,* 2970. https://doi.org/10.3389/fpsyg.2019.02970

Pace, G., Orrù, G., Monaro, M., Gnoato, F., Vitaliani, R., Boone, K. B., Sartori, G. (2019). Malingering detection of cognitive impairment with the b test is boosted using machine learning. *Frontiers in Psychology, 10*, 1650. https://doi.org/10.3389/fpsyg.2019.01650

Resnick, P. J. (1997). The malingering of posttraumatic disorders. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (2nd ed., pp. 84–103). New York, NY: Guildford Press.

Rogers, R., & Bender, S. D. (Eds.). (2018). *Clinical assessment of malingering and deception* (4th ed.). New York, NY: Guilford Press.

Rogers, R., Sewell, K. W., & Gillard, N. D. (2010). *Structured interview of reported symptoms, professional manual* (2nd ed.). Odessa, FL: Psychological Assessment Resources.

Roma, P., Giromini, L., Burla, F., Ferracuti, S., Viglione, D. J., & Mazza, C. (2019). Ecological validity of the Inventory of Problems-29 (IOP-29): An Italian study of court-ordered, psychological injury evaluations using the Structured Inventory of Malingered Symptomatology (SIMS) as a criterion variable. *Psychological Injury and Law, 13,* 57–65. https://doi.org/10.1007/s12207-019-09368-4

Roma, P., Mazza, C., Ferracuti, G., Cinti, M. E., Ferracuti, S., & Burla, F. (2019b). Drinking and driving relapse: Data from BAC and MMPI-2. *PLoS ONE*, *14*(1). doi:https://doi.org/10.1371/journal.pone.0209116

Roma, P., Mazza, C., Mammarella, S., Mantovani, B., Mandarelli, G., & Ferracuti, S. (2019c). Faking-good behavior in self-favorable scales of the MMPI-2: A study with time pressure *European Journal of Psychological Assessment,* 1–9. https://doi.org/10.1027/1015-5759/a000511

Roma, P., Piccinni, E., & Ferracuti, S. (2016). Using MMPI-2 in forensic assessment. *Rassegna Italiana di Criminologia, 10*(2), 116–122.

Roma, P., Ricci, F., Kotzalidis, G. D., Abbate, L., Lavadera, A. L., Versace, G., et al. (2014). MMPI-2 in child custody litigation: A comparison between genders. *European Journal of Psychological Assessment, 30*(2), 110–116. https://doi.org/10.1027/1015-5759/a000192

Roma, P., Verrocchio, M. C., Mazza, C., Marchetti, D., Burla, F., Cinti, M. E., & Ferracuti, S. (2018). Could time detect a faking-good attitude? A study with the MMPI-2-RF. *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.01064

Santamaría, P., Capilla Ramírez, P., & González Ordi, H. (2013). Prevalencia de simulación en incapacidad temporal: Percepción de los profesionales de la salud [Simulation prevalence in temporary disability: Perception of health professionals]. *Clínica y Salud, 24*(3), 139–151. https://doi.org/10.5093/cl2013a15

Sharf, A. J., Rogers, R., Williams, M. M., & Henry, S. A. (2017). The effectiveness of the MMPI-2-RF in detecting feigned mental

disorders and cognitive deficits: A meta analysis. *Journal of Psychopathology and Behavioral Assessment, 39,* 441–455.
Smith, G. P., & Burger, G. K. (1997). Detection of malingering: Validation of the Structured Inventory of Malingered Symptomatology (SIMS). *Journal of the American Academy of Psychiatry and the Law Online, 25*(2), 183–189. Retrieved from https://jaapl.org/
Tombaugh, T. N. (1996). *Test of memory malingering: TOMM*. New York, NY & Toronto: MHS.
Tracy, D. K., & Rix, K. J. (2017). Malingering mental disorders: Clinical assessment. *British Journal of Psychiatric Advances, 23*(1), 27–35. https://doi.org/10.1192/apt.bp.116.015958
Troldahl, V. C., & Powell, F. A. (1965). A short-form dogmatism scale for use in field studies. *Social Forces*, *44*(2), 211–214. Retrieved from https://academic.oup.com/sf
van Impelen, A., Merckelbach, H., Jelicic, M., & Merten, T. (2014). The Structured Inventory of Malingered Symptomatology (SIMS): A systematic review and meta-analysis. *The Clinical Neuropsychologist, 28*(8), 1336–1365. https://doi.org/10.1080/13854046.2014.984763
Viglione, D. J., Giromini, L., & Landis, P. (2017). The development of the Inventory of Problems–29: A brief self-administered measure for discriminating bona fide from feigned psychiatric and cognitive complaints. *Journal of Personality Assessment, 99*(5), 1–11. https://doi.org/10.1080/00223891.2016.1233882
World Health Organization (2019). International statistical classification of diseases and related health problems: 11th revision (ICD-11). Retrieved from https://icd.who.int/en
Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393
Young, G. (2014). Resource material for ethical psychological assessment of symptom and performance validity, including malingering. *Psychological Injury and Law, 7*(3), 206–235. https://doi.org/10.1007/s12207-014-9202-2
Zago, S., Piacquadio, E., Monaro, M., Orrù, G., Sampaolo, E., Difonzo, T., & Heinzl, E. (2019). The detection of malingered amnesia: An approach involving multiple strategies in a mock crime. *Frontiers in Psychiatry, 10,* 424. https://doi.org/10.3389/fpsyt.2019.00424