Schizophrenia Bulletin vol. 48 no. 2 pp. 425–436, 2022
https://doi.org/10.1093/schbul/sbab134
Advance Access publication December 16, 2021

Machine Learning Identifies Digital Phenotyping Measures Most Relevant to Negative Symptoms in Psychotic Disorders: Implications for Clinical Trials

Sayli M. Narkhede[1], Lauren Luther[1], Ian M. Raugh[1], Anna R. Knippenberg[1], Farnaz Zamani Esfahlani[2], Hiroki Sayama[3], Alex S. Cohen[4], Brian Kirkpatrick[5], and Gregory P. Strauss*[,1]

[1]Department of Psychology, University of Georgia, Athens, GA, USA; [2]Department of Psychology, Indiana University, Bloomington, IN, USA; [3]Department of Systems Science and Industrial Engineering, Binghamton University, Binghamton, NY, USA; [4]Department of Psychology, Louisiana State University, Baton Rouge, LA, USA; [5]Department of Psychiatry, University of Nevada, Reno School of Medicine, Reno, NV, USA

*To whom correspondence should be addressed; Department of Psychology, University of Georgia, 125 Baldwin St., Athens, GA 30602, USA; tel: 706-542-0307, fax: 706-542-3275, e-mail: gstrauss@uga.edu

Background: **Digital phenotyping has been proposed as a novel assessment tool for clinical trials targeting negative symptoms in psychotic disorders (PDs). However, it is unclear which digital phenotyping measurements are most appropriate for this purpose.** ***Aims:*** **Machine learning was used to address this gap in the literature and determine whether: (1) diagnostic status could be classified from digital phenotyping measures relevant to negative symptoms and (2) the 5 negative symptom domains (anhedonia, avolition, asociality, alogia, and blunted affect) were differentially classified by active and passive digital phenotyping variables.** ***Methods:*** **Participants included 52 outpatients with a PD and 55 healthy controls (CN) who completed 6 days of active (ecological momentary assessment surveys) and passive (geolocation, accelerometry) digital phenotyping data along with clinical ratings of negative symptoms.** ***Results:*** **Machine learning algorithms classifying the presence of a PD diagnosis yielded 80% accuracy for cross-validation in H_2O AutoML and 79% test accuracy in the Recursive Feature Elimination with Cross Validation feature selection model. Models classifying the presence vs absence of clinically significant elevations on each of the 5 negative symptom domains ranged in test accuracy from 73% to 91%. A few active and passive features were highly predictive of all 5 negative symptom domains; however, there were also unique predictors for each domain.** ***Conclusions:*** **These findings suggest that negative symptoms can be modeled from digital phenotyping data recorded in situ. Implications for selecting the most appropriate digital phenotyping variables for use as outcome measures in clinical trials targeting negative symptoms are discussed.**

Key words: ecological momentary assessmen t/schizophrenia/negative symptoms/machine learning/classification/feature selection/digital phenotyping

Introduction

Psychotic disorders (PDs) are chronic and associated with profound functional impairment and disability.[1–4] Negative symptoms (ie, reductions in emotion, motivation, communication, and behavior) are a strong predictor of disability and poor functioning,[5–7] highlighting their importance as a treatment target. Unfortunately, currently available pharmacological interventions have not been efficacious for remediating negative symptoms.[8]

Although contemporary clinical rating scales have allowed for important advances in understanding the phenomenology and etiology of negative symptoms,[9,10] the limitations associated with these measures, including social desirability, halo effects, and low resolution,[11–16] may make it difficult to observe robust treatment effects. Negative symptoms rating scales especially require considerable integration of both internal (ie, self-report) and behavioral symptom components across dynamic conditions such as time, environment, and activities. Failure to consider these fluctuations by using clinical rating scale items (eg, blunted affect across the past week) may lead to a misrepresentation of treatment effects. Cognitive impairments present in PDs may also make it difficult to respond to the high recall demands of negative symptom scales such as retrospectively reporting fluctuations in higher-order states like motivation and pleasure. Due to these challenges, there has been increasing interest in validating what appears to be the third generation of negative symptom measurement: digital phenotyping (ie, using mobile devices to collect data in real-life).[15,17–20] Digital phenotyping is commonly divided into active (ie, intentionally initiated by the participant) and passive (ie, unobtrusively recorded via a mobile device's background sensors) measurements.[18] Several recent studies support the feasibility of using active (eg, surveys, video recordings) and passive (eg, geolocation, accelerometry) digital phenotyping measures in PDs.[11,16,19–26] However, the psychometric evidence to date has not crossed the threshold needed for these measures to be considered viable to use in clinical trials, and it is currently unclear which digital phenotyping measures might be the most appropriate outcome measures for studies targeting negative symptoms.

This study used machine learning to address these gaps in the literature and determine whether: (1) diagnostic status could be accurately predicted from digital phenotyping measures theoretically relevant to negative symptoms (ie, identifying whether the presence of abnormality could be detected), (2) the presence (vs absence) of clinically significant elevations in the 5 negative symptom domains (anhedonia, avolition, asociality, alogia, and blunted affect) could be differentially modeled by unique active and passive digital phenotyping variables (ie, identifying whether specific measures hold relevance for individual domains or the broader negative symptom construct). A range of digital phenotyping variables hypothesized to hold relevance for negative symptom assessment were entered into machine learning algorithms to determine which combinations of active and passive variables enhanced model metrics such as accuracy, precision, recall, and the area under the receiver operating characteristic curve (ROC AUC). A comprehensive data analysis approach was utilized, which included multiple machine learning algorithms, statistical tests, and a comparative approach to determine which classification tier individual digital phenotyping variables fell into. Based on prior digital phenotyping work,[27] we hypothesized that multiple active and passive variables would be identified as key features in each evaluated model, and that models would provide cross-validation accuracies >80%.

Method

Participants

Participants included 52 outpatients with a PD and 55 healthy controls (CN). These groups were matched on age, sex, parental education, or race; however, PD had lower personal education than CN. There was a trend toward PD completing fewer EMA surveys than CN (see table 1).

PD was recruited from local community mental health centers and online or printed advertisements. Diagnoses were made using the Structured Clinical Interview for DSM-5 (SCID).[28] CN were recruited from the local community using printed and online advertisements. CN had no current SCID-5 major psychiatric diagnoses (eg, mood, substance use), no current SCID-PD[29] schizophrenia-spectrum personality disorders, no lifetime history of psychotic or bipolar disorders, no psychosis family history, and were not currently prescribed psychotropic medications. All participants reported no lifetime neurological disorders. Participants provided written informed consent for a protocol approved by the University of Georgia's Institutional Review Board.

Procedures

Study procedures occurred in 3 phases.

Phase 1 Participants completed clinical interviews and received digital phenotyping training. Diagnostic and symptom interviews were conducted by a licensed clinical psychologist or raters trained to reliability using gold-standard videos. PD outpatients were rated on the Brief Negative Symptom Scale (BNSS),[10] Positive and Negative Syndrome Scale (PANSS),[30] and Level of Functioning Scale (LOF).[31] Based on prior negative symptom theoretical conceptualizations and scale factor analyses,[32–36] we used the BNSS to measure the 5 negative symptom domains (see supplementary table S1 for correlations between domains). Digital phenotyping

Table 1. Sample Characteristics

Variable	PD (n = 52)	CN (n = 55)	Test Statistic	P
Age, *M* (SD)	38.98 (11.97)	39.07 (10.62)	$F = 0$	.966
Male, *n* (%)	18 (34.6%)	17 (30.9%)	$\chi^2 = 0.17$	.682
Personal education	13.21 (2.28)	15.4 (2.82)	$F = 19.37$	<.001
Parental education	13.83 (2.9)	13.63 (2.85)	$F = 0.12$	.731
Race			$\chi^2 = 8.62$	.125
African American	17 (32.7%)	16 (29.1%)		
Asian American	0	4 (7.3%)		
Biracial	3 (5.8%)	3 (5.5%)		
Caucasian	30 (57.7%)	24 (43.6%)		
Hispanic/Latino	2 (3.8%)	6 (10.9%)		
Other	0	2 (3.6)		
Survey adherence	57.85% (26.49%)	66.97% (23.76%)	$F = 3.52$	.063

Note: CN, control group; PD, psychotic disorders group. PD group was composed of people with schizophrenia (n = 22), schizoaffective disorder (n = 27), and bipolar disorder with psychotic features (n = 3). Adherence is the percentage of surveys completed per day, out of 8.

training involved instruction in how to use an Embrace smartband, Android smartphone provided for data collection, and mEMA app (www.ilumivu.com), including completing a practice survey in the app to ensure participant understanding of the EMA procedures.

Phase 2 Digital phenotyping data were collected for 6 days. Surveys were presented via the mEMA app randomly within 90-min epochs from 9 AM to 9 PM, aligning with prior EMA procedures.[37,38] Surveys, prompted by a tone, were available for 15 min before becoming disabled. The time interval between surveys was at least 18 min and no more than 180 min. Only apps required for study procedures (eg, mEMA) were accessible to participants.

Surveys assessed multiple psychological processes (eg, emotional experience, emotion regulation). Only those germane to negative symptoms are described here (see Supplemental Materials for complete surveys). Following prior factor analytic work[20] and BNSS procedures, surveys included items to assess both internal experience and behavioral components of anhedonia, avolition, and asociality. Behavioral components were assessed from context reports related to engagement in recreational, work/school, self-care, and social activities. Internal experience components were assessed via questions related to enjoyment, interest, and motivation for the aforementioned activities (see table 2 for items). Given the critical role of defeatist performance beliefs (DPB) as a psychological process underlying negative symptoms,[39–41] 3 items (see table 2) were included to index this construct. These items were modified to a state format from the trait DPB scale.[39] The items were developed in consultation with the DPB scale authors (AT Beck and PM Grant) who

Table 2. Ecological Momentary Assessment Survey Items and Description of Passive Digital Phenotyping Variables

Feature[a]	Items
Defeatist performance beliefs[b] (DPB)	I have to do well all the time or people will not respect me. If you cannot do something well, there is little point in doing it. If I fail at all, it is as bad as being a complete failure.
Anhedonia internal experience[b]	How much are you enjoying the activity? How much do you think you will enjoy that activity the next time you do it? How much are you enjoying this social interaction? How much do you think you will enjoy interacting with them next time?
Anhedonia behavior	What are you doing right now? Recreation.
Avolition internal experience	How interested are you in the activity?
Avolition behavior	What are you doing right now?[c] Working/Studying, Errands/Housework, Exercising, Shopping, or Commuting/Traveling.
Asociality internal experience	How interested are you in this social interaction?
Asociality behavior	Who are you interacting with?[c] Significant other, Family/Roommates, or Friends.
Geolocation (GPS) distance from home	Meters from home. Calculated using Haversine formula and Earth radius = 6 371 000.[d]
Geolocation (GPS) meters change	Meters changed between samples, using Haversine formula.[d]
Home time	Percentage of time spent within 200 m of home around each survey sample.
Accelerometry (ACL) mean	Mean total acceleration across X, Y, and Z axes.[d] Calculated as the sum of squares from each axis.
Accelerometry (ACL) standard deviation	Standard deviation of the mean acceleration across X, Y, and Z axes within the 30 min of each survey sample.
Accelerometry activity index (ACL AI)[e]	Activity index based on band accelerometry.[d]
Accelerometry Euclidean norm minus one (ACL ENMO)[e]	Band accelerometry mean minus one.[d]

Note: BNSS, Brief Negative Symptom Scale.

[a] All items are scored such that higher values indicate higher severity—negative symptom items (anhedonia, avolition, and asociality) are all reverse scored. Our categorization of negative symptoms items was based on the theoretical framework and procedures of the BNSS and a prior confirmatory factor analysis of the items supporting this scoring.[20]

[b] Average of items.

[c] If any item selected, 1, otherwise, 0.

[d] Averaged over 30 min around each survey.

[e] These items were not included in the machine learning analyses due to too much missing data.

identified these items as most critical to the DPB construct and processes underlying negative symptoms.

Passive digital phenotyping measures were collected from the smartphone (geolocation and accelerometry) and smartband (accelerometry). Table 2 contains the variables derived from these recordings.

Geolocation Geolocation data were recorded every 10 min or when participants moved more than 10 m. Participants' locations were also recorded when they completed a survey. Data were stored as GPS coordinates and changes in meters from the previous sample. Distance from home was calculated at each sample as change in GPS coordinates from the participant's home. Participant's home location was determined as the mean percentage of samples corresponding to the participant's home, as endorsed by the participant. Geolocation has demonstrated good reliability and moderate convergent validity with negative symptoms.[20,22]

Accelerometry Phone sensors were programmed to collect accelerometry with each change in *XYZ* coordinate motion (every accelerometry change being logged as a single instance), with separate value outputs for *X*, *Y*, and *Z* movement axes. The smartband collected accelerometry as gravitational force (g units) at a rate of 32 Hz (range: −16 to 16 g). Measurements of average frequency of movement, vigor of movement, and movement variability were calculated from smartband and phone accelerometry recordings using validated metrics[42] (see table 2). Accelerometry has demonstrated good reliability and moderate convergent validity with clinically rated avolition.[43]

Phase 3 Participants completed cognitive and reward processing tests not considered in this manuscript. They also returned the smartphone, smartband, chargers, and received compensation consisting of $20 per h for completing interviews and tests and $1 per survey completed. An $80 bonus was provided for returning all equipment.

Data Analysis

Data Collection Passive data (ie, geolocation, accelerometry) was averaged within the 30-min epoch around each active survey in order to pair passive and active data. Models only used complete data, where all active and passive variables of interest were present.

Machine Learning Models The same analytic strategy (see figure 1) was implemented in group classification (CN vs PD, Aim 1) and the presence (vs absence) of the 5 negative symptoms (Aim 2). Machine learning algorithms for both Aims explored 12 digital phenotyping features relevant to negative symptoms (see table 2), with Aim 1 having 2340 data points and Aim 2 having 965 data points. There are fewer data points for Aim 2 because CN did not complete the BNSS. Six sets of supervised machine learning algorithms were examined, 1 set for Aim 1 and 5 sets for Aim 2 (1 for each negative symptom), with each set comprising the same 6 feature selection methods and 3 statistical tests. Model performance for identified features was measured using 3 classifiers. The presence of a negative symptom was determined by a score of 2 (mild) or higher on any BNSS item within a given domain: anhedonia (present/absent participants = 30/23), avolition (33/2), asociality (28/25), blunted affect (27/26), and alogia (7/46).

Feature Selection To examine which features were most consistently identified, we used a comprehensive feature selection approach consisting of 6 machine learning methods (Boruta,[44] Recursive Feature Elimination with Cross Validation (RFECV), Logistic Regression using Statsmodels,[45] Random Forest, H_2O,[46] and L1 Regularization) and 3 statistical tests (Chi Square, Kendall's Rank Coefficient Method, and Select Percentile with ANOVA *F*-value). To identify feature importance in H_2O AutoML, we used H_2O XGBoost (eXtreme Gradient Boosting) and H_2O GBM (Gradient Boosting Machine), which provided feature importance based on observed cross-validation accuracy and AUC.

Steps were taken to determine feature tiers. First, top tier features were identified by examining which features were ranked as most important in each feature selection method. To be determined top tier, a feature was required to be selected as the highest rank across most feature selection methods and be statistically significant in at least 1 statistical test. Features in the second and third tiers were identified by consistency of rankings across methods, selection frequency, and statistical test significance; however, feature rankings were allowed to vary across feature selection methods. Features identified as second tier were selected more across ML methods than third tier features, which were mostly observed to be statistically significant and identified rarely across the feature selection methods.

Machine Learning Training and Evaluation Overview We split the sample into a training and cross-validation dataset and test dataset. In both datasets, model performance for subsets of consistently selected features was estimated using 3 learning classifiers (Random Forest, *K*-Nearest Neighbors [KNN], and Logistic Regression [this served as the baseline comparison model]).

Training and Cross-validation Dataset Seventy-five percent of the data was used for training and validation. We optimized classifiers by implementing Stratified K-folds Cross-validations and performing the RandomizedSearchCV technique to discover the optimal hyperparameters for all sets of classifiers in both aims. These procedures allow for an unbiased evaluation of model fit.

Test Dataset The remaining 25% of data was used to evaluate tuned classifiers' performance. Five classification

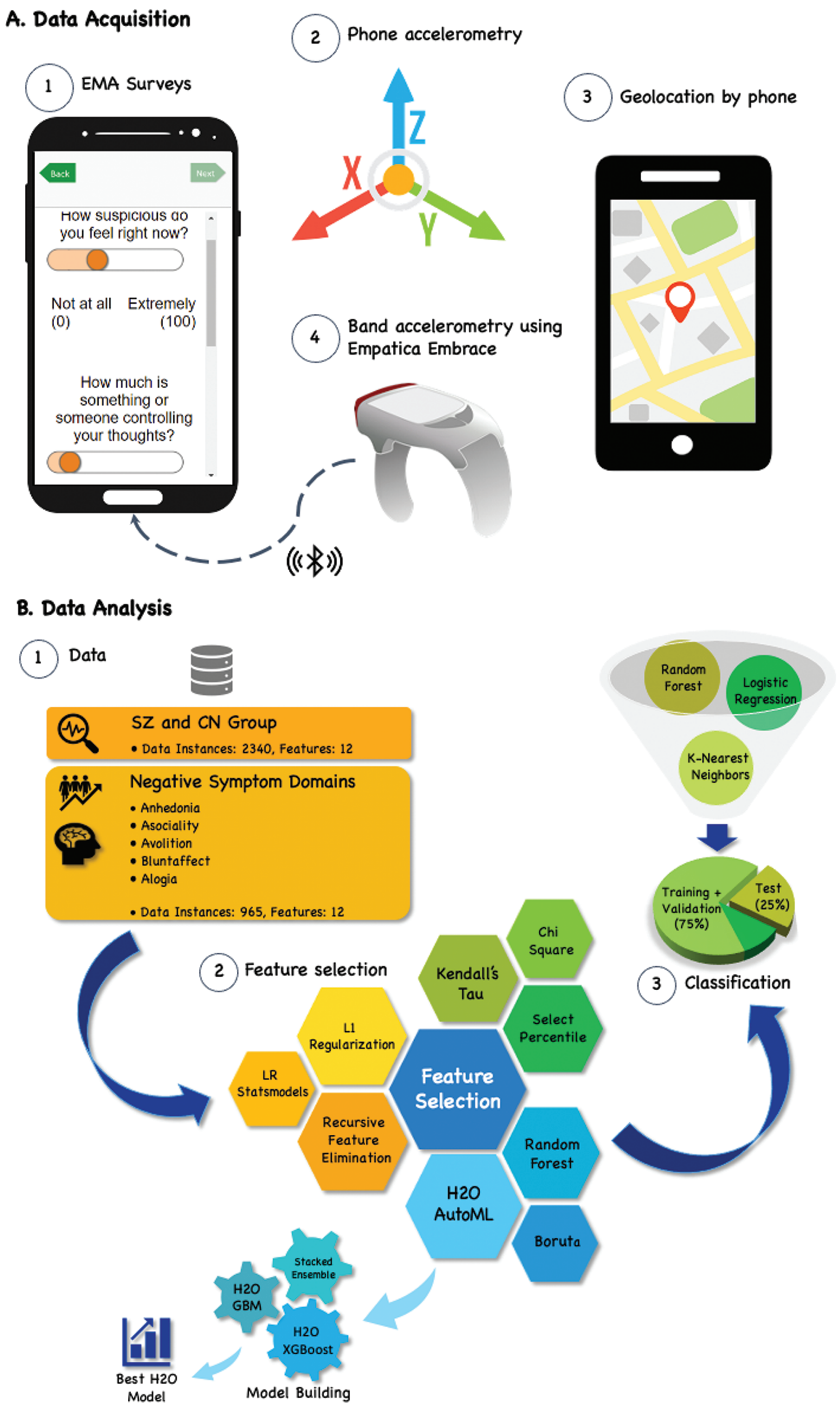

Fig. 1. Digital Phenotyping Data Acquisition and Machine Learning Analysis Pipeline

performance metrics were used to evaluate the fit of classifiers: (1) accuracy (most commonly used metric for evaluating machine learning models to check how often a classifier is correct), (2) precision (percentage of positive instances out of total predicted positive instances), (3) recall (or sensitivity; gives a true positive rate or a ratio of true positives to total actual positive instances), (4) ROC AUC prediction scores, and (5) confusion matrix (a summary table used to describe a classifier's predictive results). Guided by prior work,[27,47] values ≥.7 were used to indicate adequate fit for all classification performance metrics except for the confusion matrix, which was used to check a classifier's prediction and error rates.

In both aims, the 5 classification performance metrics were used to: (1) identify the best fitting classifier and if KNN and Random Forest outperformed the baseline classifier model across 5 feature selection methods (H_2O excluded), (2) determine ranks of feature importance by observing all model results, (3) examine how feature ranking changed when compared with the best model, and (4) check if the top tier feature remained the same across all feature selection methods and if new feature/s were included.

Model Comparison We completed several steps to have the greatest likelihood of identifying key features relevant to predicting diagnosis and the presence of negative symptom domains since individual models may produce varying results. To confirm that top tier features were indeed important features, classifier performance was also evaluated using only the top features. Performance metrics of these classifiers (with top feature/s only) were compared with the results of classifiers (with all relevant features), and comparable performance statistics were taken as indication that the top tier features were indeed important. Consistency across models was deemed to reflect a more stable estimate of feature importance.

Results

Digital Phenotyping Adherence

Survey adherence and passive data rates differed across diagnostic and negative symptom groups. PD and negative symptom present subgroups had less smartband accelerometry data than CN and PD without negative symptoms (see supplementary table S2).

Autocorrelations of the digital phenotyping measures demonstrated expected variability of these measures across the data collection period (see supplementary table S3).

Aim 1: Machine Learning Classification of Diagnostic Status

Feature selection methods identified a range of 2 (Random Forest) to 7 (L1 Regularization) relevant features in PD and CN groups. Random Forest classifier in RFECV method performed relatively well on test data when compared with the performance measures of other models. Random Forest classifier outperformed both KNN and Logistic Regression with above adequate performance metrics. H_2O XGBoost delivered the highest cross-validation accuracy and AUC (≥0.80) (see table 3).

The top tier/key features predicting PD diagnosis were DPB and geolocation home distance mean. Second tier features included avolitional behavior, anhedonia internal experience, asociality internal experience, and avolitional internal experience. The third tier feature was phone accelerometry standard deviation. Phone accelerometry mean did not meet any tier selection criteria (see table 4). Follow-up models entering only the top tier features confirmed the importance of the top tier variables, as performance statistics did not change appreciably from the original model (see supplementary table S4).

Aim 2: Machine Learning Classification of the Presence of Individual Negative Symptoms

Feature selection methods for negative symptoms ranged from 4 (Random Forest) to 8 (Logistic Regression) features in anhedonia, 3 (Logistic Regression) to 9 (RFECV) features in asociality, 2 (RFECV) to 7 (Random Forest) features in avolition, 3 (RFECV) to 8 (Boruta, Logistic Regression, L1 Regularization) features in blunted affect, and 4 (RFECV) to 8 (Boruta) features in alogia (see supplementary table S5). Random Forest provided the best performance metrics amongst the 3 classifiers. Random Forest and KNN classifiers had better classification performance metrics than Logistic Regression across all models. H_2O XGBoost or H_2O GBM in H_2O AutoML delivered the highest cross-validation metrics for both accuracy and AUC for each domain. The highest classification performance metrics were estimated on test data by Random Forest classifier in Random Forest feature selection method in anhedonia, Random Forest classifier in RFECV and L1 Regularization in asociality, Random Forest classifier in Boruta in avolition, Random Forest classifier in RFECV in blunted affect, and Random Forest classifier in Boruta and Random Forest feature selection methods in alogia. All of the best fitting BNSS domain models had adequate performance metrics (except for recall in alogia).

To identify the optimal features, consistency of feature appearance was compared across all feature selection methods and statistical tests for each BNSS domain (see table 5). The optimal features for domains were:

1. Anhedonia: top tier features were DPB and anhedonia internal experience; second tier features were geolocation home distance standard deviation, geolocation home distance mean, phone accelerometry standard deviation, phone accelerometry mean, and avolition internal experience; and third tier features were asociality internal experience.
2. Asociality: top tier features were phone accelerometry standard deviation, geolocation home distance mean, and DPB; second tier features were anhedonia internal experience, and phone accelerometry mean; and third tier features were geolocation home distance standard deviation, avolition internal experience, asociality internal experience, and geolocation meters change mean.

Table 3. Machine Learning Results in Classifying Schizophrenia Diagnosis

			Performance Measures			
Feature Selection Models	Selected Features		Classifiers	RF	LR	KNN
Boruta	DPB		Accuracy	**0.769**	0.697	0.723
	GPS home distance mean		Precision	**0.742**	0.694	0.683
	ACL mean		Recall	**0.703**	0.518	0.627
	ACL SD		ROC_AUC	**0.843**	0.727	0.789
	Anhedonia internal experience		Confusion	**[275 61]**	[279 57]	[270 71]
	Avolition internal experience		matrix	**[74 175]**	[120 129]	[91 153]
	GPS home distance SD (occasionally identified)					
	Asociality internal experience (occasionally identified)					
RFECV	DPB		Accuracy	**0.79**	0.696	0.72
	GPS home distance mean		Precision	**0.774**	0.69	0.684
	GPS home distance SD		Recall	**0.715**	0.518	0.635
	Anhedonia internal experience		ROC_AUC	**0.86**	0.725	0.79
	Avolition internal experience		Confusion	**[284 52]**	[278 58]	[263 73]
	Asociality internal experience		matrix	**[71 178]**	[120 129]	[91 158]
Logistic Regression using	DPB		Accuracy	**0.771**	0.697	0.711
Statsmodels ($P < .05$)	Avolition behavior		Precision	**0.745**	0.696	0.68
	GPS home distance mean		Recall	**0.703**	0.514	0.624
	ACL mean		ROC_AUC	**0.84**	0.727	0.771
	Asociality internal experience		Confusion	**[276 60]**	[280 56]	[260 75]
	Anhedonia internal experience ($P = .0508$)		matrix	**[74 175]**	[121 128]	[94 156]
	Features	Feature Importance	Classifiers	RF	LR	KNN
Random Forest	DPB[a]	0.277	Accuracy	**0.747**	0.699	0.718
	GPS home distance mean[a]	0.163	Precision	**0.693**	0.695	0.691
	Anhedonia internal experience	0.076	Recall	**0.732**	0.522	0.616
	Asociality internal experience	0.071	ROC_AUC	**0.818**	0.727	0.776
	Avolition internal experience	0.07	Confusion	**[254 81]**	[279 57]	[266 69]
	ACL SD	0.066	matrix	**[67 183]**	[119 130]	[96 154]
	GPS home distance SD	0.065				
	ACL mean	0.065				
	GPS meters change SD	0.059				
	GPS meters change mean	0.058				
	Avolition behavior	0.018				
	Asociality behavior	0.012				
	Features	Scaled Importance	Classifier	H_2O XGBoost		
H_2O AutoML	DPB	1.000	Accuracy	**0.796**		
	GPS home distance mean	0.562	AUC	**0.864**		
	Anhedonia internal experience	0.329				
	Asociality internal experience	0.275				
	Avolition internal experience	0.234				
	ACL SD	0.215				
	ACL mean	0.214				
	GPS home distance SD	0.116				
	Avolition behavior	0.074				

Table 3. Continued

	Features	Scaled Importance	Classifier	H_2O XGBoost		
	Asociality behavior	0.054				
	GPS meters change mean	0.053				
	GPS meters change SD	0.048				
	Features	Estimated Coefficients	Classifiers	RF	LR	KNN
L1 Regularization	DPB[a]	0.805	Accuracy	**0.762**	0.696	0.723
	Avolition behavior[a]	−0.409	Precision	**0.727**	0.692	0.712
	GPS home distance mean[a]	−0.134	Recall	**0.712**	0.514	0.59
	Asociality internal experience[a]	−0.06	ROC_AUC	**0.845**	0.728	0.766
	Anhedonia internal experience[a]	0.032	Confusion matrix	**[268 67]** **[72 178]**	[279 57] [121 128]	[277 59] [103 146]
	Avolition internal experience[a]	0.016				
	ACL SD[a]	−0.001				

The value counts of classes ie, control and presence of group are 1339 and 1001. Best classifier(s) for model is bolded. *Note*: ACL, accelerometry; DPB, defeatist performance beliefs; KNN, *K*-Nearest Neighbors; LR, Logistic Regression; RF, Random Forest; ROC_AUC, area under the receiver operating characteristic curve.

[a] Features were considered for estimating final fit of models.

Table 4. Statistical Tests Results for Identifying Features Significance in Schizophrenia Diagnosis

Kendall's Tau		Select Percentile With ANOVA *F*-Value	
Selected Features	P	Selected Features	P
DPB	<.001	DPB	<.001
Avolition behavior	<.001	Avolition behavior	<.001
GPS home distance mean	<.001	GPS home distance mean	<.001
Anhedonia internal experience	<.001	Avolition internal experience	<.001
Avolition internal experience	<.001	Anhedonia internal experience	<.001
ACL SD	<.05	ACL SD	<.05

The Chi Square test to determine statistical significance of categorical features only selected avolition behavior. Only ranking of anhedonia internal experience and avolition internal experience interchanged in statistical tests; rest of the features remained unchanged. *Note*: ACL, accelerometry; DPB, defeatist performance beliefs.

3. Avolition: top tier features were geolocation home distance mean and DPB; second tier feature was phone accelerometry mean; and third tier features were asociality internal experience, phone accelerometry standard deviation, and anhedonia internal experience. Geolocation home distance standard deviation was rarely identified.
4. Blunted affect: top tier features were DPB and geolocation home distance mean; second tier features were geolocation meters change mean, phone accelerometry mean, phone accelerometry standard deviation, anhedonia internal experience, and avolition internal experience; and third tier features were asociality behavior and geolocation home distance standard deviation.
5. Alogia: top tier features were DPB and geolocation home distance mean; second tier features were anhedonia internal experience, asociality internal experience, and avolition internal experience; and third tier features were avolition behavior and phone accelerometry mean.

Follow-up models entering only top tier features confirmed the importance of these features for each negative symptom domain since performance statistics were comparable to the original model (see supplementary table S6).

Discussion

This study used machine learning to determine which digital phenotyping variables are most relevant for classifying diagnostic status and the presence of clinically significant elevations in each of the 5 negative symptom domains: anhedonia, avolition, asociality, blunted affect, and alogia. A comprehensive machine learning approach was adopted to evaluate consistency of features identified and model performance across 6 feature selection models and 3 statistical tests. This approach allows for greater assurance that conclusions are not biased by particular models implemented. Several findings emerged that have important implications for selecting digital phenotyping measurements for negative symptom assessment.

Table 5. Statistical Tests Results for Identifying Features Significance in Negative Symptom Domains

Feature Selection Methods	Negative Symptom Domains				
	Anhedonia	Asociality	Avolition	Blunted Affect	Alogia
Kendall's Tau ($P < .05$)	DPB Anhedonia internal experience Avolition internal experience Asociality internal experience	ACL SD GPS home distance mean	GPS home distance mean ACL mean Anhedonia internal experience ACL SD Asociality internal experience DPB	DPB Asociality behavior Anhedonia internal experience ACL SD Avolition internal experience ACL mean Asociality internal experience	Anhedonia internal experience Asociality internal experience DPB Avolition internal experience GPS home distance mean Asociality behavior GPS home distance SD GPS meters change SD GPS meters change mean Avolition behavior
Select Percentile AVONA *F*-value ($P < .05$)	DPB Anhedonia internal experience Avolition internal experience GPS home distance SD Asociality internal experience	ACL SD GPS home distance mean	GPS home distance mean ACL mean ACL SD Anhedonia internal experience Asociality internal experience	DPB Asociality behavior ACL SD GPS home distance SD Anhedonia internal experience Avolition internal experience GPS home distance mean GPS meters change SD ACL mean GPS meters change mean Asociality internal experience	Anhedonia internal experience Asociality internal experience DPB Avolition internal experience Asociality behavior GPS home distance mean Avolition behavior

Note: ACL, accelerometry; DPB, defeatist performance beliefs. Chi Square test which was performed to determine statistical significance of categorical features, selected asociality behavior in blunted affect domain and both, avolition behavior and asociality behavior in alogia domain. In asociality, DPB was not statistically significant ($P > .05$). However, this feature was identified across most of the ML models and performance of random forest increased significantly after assessing metrics with DPB in top tier. Thus, it was included as a top tier feature.

First, a PD diagnosis was predicted with cross-validation accuracy at ~80% and ~79% test accuracy using a combination of active and passive digital phenotyping measures. These rates are comparable to what has been observed in most other serious mental illness digital phenotyping machine learning studies.[27] These findings suggest that as a class of assessment tools, digital phenotyping is capable of detecting the presence of abnormalities in behaviors and experiences characteristic of negative symptoms. Knowing that these measures can detect deficit states compared to a healthy group is an important first step before they can then be applied in studies only using clinical groups, such as pharmaceutical trials. Given that both active and passive measures were identified as key features suggest that a combination of both data streams may be optimal. This level of accuracy in diagnostic prediction is noteworthy given that only measures putatively relevant for negative symptoms were included, and negative symptoms can be observed in the general population; assessments of positive and disorganized symptoms, which are less common in the general population, were intentionally not incorporated to allow the hypotheses of interest to be tested. However, inclusion of such measures would likely further enhance model prediction.

Second, models classifying the presence of each of the 5 negative symptom domains varied in accuracy, precision, recall, ROC AUC, and confusion matrix. However, test accuracy was between 73% and 91% for the best models of individual domains. Observation of model performance at this level for blunted affect and alogia was surprising given that the variables examined (geolocation, accelerometry, and EMA surveys) primarily align with anhedonia, avolition, and asociality, and there were no measures included that had face validity for measuring alogia or blunted affect. The model performance achieved by alogia may reflect that this symptom is often a marker of illness severity and is typically only present in those with the most severe psychopathology. Notably, there were some variables that were predictive of all 5 domains, including DPB and geolocation home distance. The observation that defeatist beliefs were a top feature for each domain supports prior evidence that this construct is an important psychological mechanism of negative symptoms, broadly defined.[39–41] The relative importance of geolocation home distance across multiple negative symptom domains suggests that it may be the best global marker of negative symptoms among the passive variables. Both mean and standard deviation scores for home distance may also be worth computing. For anhedonia, avolition, and asociality, the *experiential* EMA items generally appeared in the higher tiers among selected features, suggesting that these items may hold value for assessing their intended constructs. However, the *behavioral* EMA survey items for these 3 domains were less likely to be selected as critical features. These behavioral items were taken from participant context reports (ie, who they were with [social], what they were doing [activity and location]). It is possible that this approach to assessing the behavioral components of these domains is not as robust as we anticipated. It may be more ideal to combine the experiential EMA items with objective behavioral markers from the passively recorded accelerometry or geolocation data than with behavioral EMA items. However, context may still be important to collect in EMA surveys, as it allows for a richer understanding of the passive data (eg, looking at geolocation only in certain contexts such as when participants are outside of their home).

Third, there may be unique combinations of active EMA surveys and passive measures that are optimal for anhedonia, avolition, and asociality. Based on the machine learning findings observed here and the face validity of these assessments, we have the following recommendations:

1. For the avolition domain, the combination of avolition internal experience, geolocation home distance, and phone accelerometry mean measures may be ideal.
2. The asociality domain may be best assessed via a combination of geolocation home distance, phone accelerometry standard deviation, and asociality experiential EMA surveys.
3. The anhedonia domain may be best assessed by geolocation home distance mean and standard deviation, phone accelerometry mean and standard deviation, and anhedonia experiential EMA surveys measuring consummatory and anticipatory pleasure.
4. Given that anhedonia, avolition, and asociality each have both experiential and behavioral components,[36] combining subjective and objective digital phenotyping measurements of negative symptoms may be necessary to assess the constructs as they are currently operationalized.

These findings can also inform measurement in negative symptom experimental psychopathology and clinical trial research. The comprehensive machine learning approach used here identified combinations of active and passive digital phenotyping variables relevant for predicting the presence of negative symptoms. The variables included in this study were not ideal for measuring alogia or blunted affect; however, measures such as ambulatory videos[19] and passive vocal recording[48] may hold promise for those constructs. Additional measures, especially those with less participant burden than EMA, may also be worth exploring for avolition, anhedonia, and asociality, such as social media data, typing keystrokes, passive vocal recording, linguistic text-message analysis, and ambulatory psychophysiology. Similarly, work is needed to more comprehensively examine the utility of adding within-person variability of these digital measures as negative symptoms markers. If negative symptoms can be quantified in real-world environments using digital phenotyping measurements, negative symptom studies might be able to be siteless in the future. This would be an enormous benefit in terms of practicality, cost, and participant burden. To determine whether digital phenotyping studies are ready for use specifically in clinical trials, larger-scale psychometric studies are needed to evaluate the reliability, validity, and sensitivity to change of a variety of these tools. There will undoubtedly be issues to work through, such as the appropriate level of resolution for each measure, how to pair active and passive variables, how to use survey context to inform passive measurement, ways to promote adherence, identifying the most appropriate analytic techniques and models, and data acquisition, storage, and confidentiality. Such studies will be necessary to overcome limitations of the current study (eg, modest sample size, lack of a replication sample, limited range of passive measures, measurements focused on 3/5 domains), which provides preliminary evidence suggesting that these digital phenotyping tools have promise for use in clinical trials.

Supplementary Material

Supplementary material is available at *Schizophrenia Bulletin.*

Funding

Research was supported by R21-MH112925 to G.P.S. from the National Institute of Mental Health. Study sponsors had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Acknowledgments

G.P.S., B.K., and A.S.C. are founders of Quantic Innovations and receive royalties for consultation and services provided related to that company; those services were not involved with the current research, and the data and mobile survey used here will not be directly used by

Quantic Innovations. However, the company may reference this article. G.P.S. and B.K. are codevelopers of the BNSS and receive fees in conjunction with its use, which are donated to the Brain & Behavior Research Foundation.

References

1. Haro JM, Novick D, Suarez D, Alonso J, Lépine JP, Ratcliffe M; SOHO Study Group. Remission and relapse in the outpatient care of schizophrenia: three-year results from the Schizophrenia Outpatient Health Outcomes study. *J Clin Psychopharmacol.* 2006;26(6):571–578.
2. Harvey PD, Heaton RK, Carpenter WT Jr, Green MF, Gold JM, Schoenbaum M. Functional impairment in people with schizophrenia: focus on employability and eligibility for disability compensation. *Schizophr Res.* 2012;140(1–3):1–8.
3. Haro JM, Novick D, Bertsch J, Karagianis J, Dossenbach M, Jones PB. Cross-national clinical and functional remission rates: Worldwide Schizophrenia Outpatient Health Outcomes (W-SOHO) study. *Br J Psychiatry.* 2011;199(3):194–201.
4. Eack SM, Newhill CE. Psychiatric symptoms and quality of life in schizophrenia: a meta-analysis. *Schizophr Bull.* 2007;33(5):1225–1237.
5. Foussias G, Mann S, Zakzanis KK, van Reekum R, Agid O, Remington G. Prediction of longitudinal functional outcomes in schizophrenia: the impact of baseline motivational deficits. *Schizophr Res.* 2011;132(1):24–27.
6. Alessandrini M, Lançon C, Fond G, et al. A structural equation modelling approach to explore the determinants of quality of life in schizophrenia. *Schizophr Res.* 2016;171(1–3):27–34.
7. Fervaha G, Foussias G, Agid O, Remington G. Impact of primary negative symptoms on functional outcomes in schizophrenia. *Eur Psychiatry.* 2014;29(7):449–455.
8. Fusar-Poli P, Papanastasiou E, Stahl D, et al. Treatments of negative symptoms in schizophrenia: meta-analysis of 168 randomized placebo-controlled trials. *Schizophr Bull.* 2015;41(4):892–899.
9. Kring AM, Gur RE, Blanchard JJ, Horan WP, Reise SP. The Clinical Assessment Interview for Negative Symptoms (CAINS): final development and validation. *Am J Psychiatry.* 2013;170(2):165–172.
10. Kirkpatrick B, Strauss GP, Nguyen L, et al. The brief negative symptom scale: psychometric properties. *Schizophr Bull.* 2011;37(2):300–305.
11. Cohen AS, Schwartz E, Le TP, et al. Digital phenotyping of negative symptoms: the relationship to clinician ratings. *Schizophr Bull.* 2021;47(1):44–53.
12. Cohen AS, Schwartz E, Le TP, Fedechko T, Kirkpatrick B, Strauss GP. Using biobehavioral technologies to effectively advance research on negative symptoms. *World Psychiatry.* 2019;18(1):103–104.
13. Cohen AS, Schwartz E, Le T, et al. Validating digital phenotyping technologies for clinical use: the critical importance of "resolution". *World Psychiatry.* 2020;19(1):114–115.
14. Ben-Zeev D, McHugo GJ, Xie H, Dobbins K, Young MA. Comparing retrospective reports to real-time/real-place mobile assessments in individuals with schizophrenia and a nonclinical comparison group. *Schizophr Bull.* 2012;38(3):396–404.
15. Torous J, Keshavan M. A new window into psychosis: the rise digital phenotyping, smartphone assessment, and mobile monitoring. *Schizophr Res.* 2018;197:67–68.
16. Mote J, Fulford D. Ecological momentary assessment of everyday social experiences of people with schizophrenia: a systematic review. *Schizophr Res.* 2020;216:56–68.
17. Insel TR. Digital phenotyping: technology for a new science of behavior. *JAMA.* 2017;318(13):1215–1216.
18. Onnela JP, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology.* 2016;41(7):1691–1696.
19. Cohen AS, Cowan T, Le TP, et al. Ambulatory digital phenotyping of blunted affect and alogia using objective facial and vocal analysis: proof of concept. *Schizophr Res.* 2020;220:141–146.
20. Raugh IM, James SH, Gonzalez CM, et al. Geolocation as a digital phenotyping measure of negative symptoms and functional outcome. *Schizophr Bull.* 2020;46(6):1596–1607.
21. Torous J, Keshavan M, Onnela JP, Staples P, Barnett I. M48. Digital phenotyping in schizophrenia using smartphones. *Schizophr Bull.* 2017;43(suppl 1):S228.
22. Depp CA, Bashem J, Moore RC, et al. GPS mobility as a digital biomarker of negative symptoms in schizophrenia: a case control study. *NPJ Digit Med.* 2019;108(2).
23. Granholm E, Holden JL, Mikhael T, et al. What do people with schizophrenia do all day? Ecological momentary assessment of real-world functioning in schizophrenia. *Schizophr Bull.* 2020;46(2):242–251.
24. Johnson EI, Grondin O, Barrault M, et al. Computerized ambulatory monitoring in psychiatry: a multi-site collaborative study of acceptability, compliance, and reactivity. *Int J Methods Psychiatr Res.* 2009;18(1):48–57.
25. Ben-Zeev D, Wang R, Abdullah S, et al. Mobile behavioral sensing for outpatients and inpatients with schizophrenia. *Psychiatr Serv.* 2016;67(5):558–561.
26. Wang R, Scherer EA, Walsh M, et al. Predicting symptom trajectories of schizophrenia using mobile sensing. *Proc ACM Interact Mob Wearable Ubiquitous Technol.* 2017;1(3):1–24.
27. Benoit J, Onyeaka H, Keshavan M, Torous J. Systematic review of digital phenotyping and machine learning in psychosis spectrum illnesses. *Harv Rev Psychiatry.* 2020;28(5):296–304.
28. First MB, Williams JBW, Karg RS, Spitzer RL. *Structured Clinical Interview for DSM-5—Research Version (SCID-5 for DSM-5, Research Version; SCID-5-RV).* Washington, DC: American Psychiatric Association; 2015.
29. First MB, Spitzer RL, Williams JBW. *Structured Clinical Interview for DSM-5(R) Personality Disorders (SCID-5-PD).* 1st ed. Washington, DC: American Psychiatric Publishing; 2015:52.
30. Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophr Bull.* 1987;13(2):261–276.
31. Hawk AB, Carpenter WT Jr, Strauss JS. Diagnostic criteria and five-year outcome in schizophrenia. A report from the International Pilot Study of schizophrenia. *Arch Gen Psychiatry.* 1975;32(3):343–347.
32. Mucci A, Vignapiano A, Bitter I, et al. A large European, multicenter, multinational validation study of the Brief Negative Symptom Scale. *Eur Neuropsychopharmacol.* 2019;29(8):947–959.

33. Ahmed AO, Kirkpatrick B, Galderisi S, et al. Cross-cultural validation of the 5-factor structure of negative symptoms in schizophrenia. *Schizophr Bull.* 2019;45(2):305–314.
34. Strauss GP, Nuñez A, Ahmed AO, et al. The latent structure of negative symptoms in schizophrenia. *JAMA Psychiatry.* 2018;75(12):1271–1279.
35. Strauss GP, Ahmed AO, Young JW, Kirkpatrick B. Reconsidering the latent structure of negative symptoms in schizophrenia: a review of evidence supporting the 5 consensus domains. *Schizophr Bull.* 2019;45(4):725–729.
36. Kirkpatrick B, Fenton WS, Carpenter WT Jr, Marder SR. The NIMH-MATRICS consensus statement on negative symptoms. *Schizophr Bull.* 2006;32(2):214–219.
37. Gard DE, Sanchez AH, Cooper K, Fisher M, Garrett C, Vinogradov S. Do people with schizophrenia have difficulty anticipating pleasure, engaging in effortful behavior, or both? *J Abnorm Psychol.* 2014;123(4):771–782.
38. Granholm E, Loh C, Swendsen J. Feasibility and validity of computerized ecological momentary assessment in schizophrenia. *Schizophr Bull.* 2008;34(3):507–514.
39. Grant PM, Beck AT. Defeatist beliefs as a mediator of cognitive impairment, negative symptoms, and functioning in schizophrenia. *Schizophr Bull.* 2009;35(4):798–806.
40. Campellone TR, Sanchez AH, Kring AM. Defeatist performance beliefs, negative symptoms, and functional outcome in schizophrenia: a meta-analytic review. *Schizophr Bull.* 2016;42(6):1343–1352.
41. Granholm E, Holden J, Worley M. Improvement in negative symptoms and functioning in cognitive-behavioral social skills training for schizophrenia: mediation by defeatist performance attitudes and asocial beliefs. *Schizophr Bull.* 2018;44(3):653–661.
42. Bai J, Di C, Xiao L, et al. An activity index for raw accelerometry data and its comparison with other activity metrics. *PLoS One.* 2016;11(8):e0160644.
43. Strauss GP, Raugh IM, Luther L, et al. Validation of accelerometry as a digital phenotyping measure of negative symptoms in schizophrenia.
44. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw.* 2010;36(11):1–13.
45. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with Python. In: Proceedings of the 9th Python in Science Conference. SciPy; 2010:92–96.
46. Candel A, Parmar V, LeDell E, Arora A. *Deep Learning with H_2O.* Mountain View, CA: H2O ai Inc; 2016.
47. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol.* 2010;5(9):1315–1316.
48. Cohen AS, Fedechko TL, Schwartz EK, et al. Ambulatory vocal acoustics, temporal dynamics, and serious mental illness. *J Abnorm Psychol.* 2019;128(2):97–105.