American College of Neuropsychopharmacology

www.nature.com/npp

ARTICLE

Check for updates

AI-based analysis of social media language predicts addiction treatment dropout at 90 days

Brenda Curtis[1✉], Salvatore Giorgi[1,2], Lyle Ungar[2,3], Huy Vu[4], David Yaden[3,5], Tingting Liu[1,3], Kenna Yadeta[1] and H. Andrew Schwartz[4]

The reoccurrence of use (relapse) and treatment dropout is frequently observed in substance use disorder (SUD) treatment. In the current paper, we evaluated the predictive capability of an AI-based digital phenotype using the social media language of patients receiving treatment for substance use disorders ($N = 269$). We found that language phenotypes outperformed a standard intake psychometric assessment scale when predicting patients' 90-day treatment outcomes. We also use a modern deep learning-based AI model, Bidirectional Encoder Representations from Transformers (BERT) to generate risk scores using pre-treatment digital phenotype and intake clinic data to predict dropout probabilities. Nearly all individuals labeled as low-risk remained in treatment while those identified as high-risk dropped out (risk score for dropout AUC = 0.81; $p < 0.001$). The current study suggests the possibility of utilizing social media digital phenotypes as a new tool for intake risk assessment to identify individuals most at risk of treatment dropout and relapse.

Neuropsychopharmacology; https://doi.org/10.1038/s41386-023-01585-5

INTRODUCTION

In the treatment of substance use disorders (SUD), reoccurrence of use (relapse) and treatment dropout remain the norm rather than the exception [1–3]. Attempts to improve treatment involve assessing the broad spectrum of heterogenous factors that affect relapse: addiction severity, demographics, life experiences, and risk of negative outcomes. Such factors are assessed during intake to treatment programs primarily via structured interviews, carried out by a trained counselor. Considered the gold standard of assessment, these interviews have enhanced treatment outcomes [4–6], but their reliance on retrospective self-report of patients often leaves an *ecological information gap* as compared to the highly complex and heterogeneous way in which SUD develops in each individual's daily life over the years before intake.

Recently, the use of social media has become nearly ubiquitous with most adults across the world using it at least once daily [7], including patients receiving SUD treatment [8, 9]. Social media posts are written in real-time, ecologically as life happens, but remain available for analysis years later [10]. Computational tools from artificial intelligence (AI) have recently been used to produce so-called *digital phenotypes*–quantitative characterizations of an individual's digital behavior–that have been shown to capture ecological and psychological factors from social media language [5, 11, 12]. Among others, language from Twitter or Facebook posts has been used to predict the presence of a depression diagnosis [13], capture personality scores as accurately as one's friends [14, 15], and predict US county rates of excessive drinking beyond demographic and socioeconomic measurements [16]. Most recently, an artificial intelligence innovation in how to do text processing, named *transformers*, has led to state-of-the-art results in such mental health predictive tasks [11, 17]. However, such techniques have yet to be applied and evaluated for assessing the risk of negative treatment outcomes for individuals in a clinical treatment setting.

Here, we evaluate the ability of a digital phenotype to help fill the *ecological information gap* during SUD intake. We create a digital phenotype from patients' Facebook posts, prior to intake, utilizing a modern deep learning-based AI model, Bidirectional Encoder Representations from Transformers (BERT) [18]. BERT produces contextualized word representations–a quantitative encoding of a word's meaning in context–and has, to date, had limited use to predict outcomes of medical interest [17, 19]. Since its development, BERT and equivalent Transformer models have transformed the way modern AI language analyses are done and are used in nearly all applications (e.g., Google's and Bing's search and question answering, Amazon Alexa, Google Voice, Apple Siri, as well as nearly all automatic Translation programs released in the past couple years) [20], but had yet to be applied to this application.

Applying the transformer BERT in order to encode a digital phenotype, our primary research questions include: (1) *How accurate is the digital phenotype for predicting future SUD treatment outcomes as compared to the Addiction Severity Index, a widely used structured intake interview?* and (2) *Which categorization of treatment outcomes is best predicted by the digital phenotype – one defining relapse broadly as anyone reporting it or one that divides relapse based on whether the patient remained in treatment or not?* To the best of our knowledge, this is the first study to apply

[1]Intramural Research Program, National Institute on Drug Abuse, Baltimore, MD, USA. [2]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA. [3]Positive Psychology Center, University of Pennsylvania, Philadelphia, PA, USA. [4]Department of Computer Science, Stony Brook University, Stony Brook, NY, USA. [5]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA. ✉email: brenda.curtis@nih.gov

Received: 4 November 2022 Revised: 3 April 2023 Accepted: 5 April 2023
Published online: 24 April 2023

and evaluate social media language for assessing the risk of negative SUD treatment outcomes over individuals within a clinical treatment setting.

MATERIALS AND METHODS

Overview of study design

Figure 1A depicts our study design. Broadly, we gathered Facebook posts from a cohort of 504 consenting patients attending outpatient SUD treatment in Philadelphia. Of these, just over half ($N = 269$) had sufficient language data—200 words, a threshold where previous language-based mental health prediction accuracy has tended to stabilize [10]. Those who met this criterion did not differ significantly in demographics or in drug history severity, while the overall sample was majority male and African American (Table 1). This was also true for missingness: there was no relationship between treatment outcomes and insufficient language data. The digital phenotype was composed of a dimension-reduced average of BERT embeddings that has recently been benchmarked to show effectiveness for a variety of language-based psychological assessments [21]. Following the prediction of clinically diagnosed depression in [13], the digital phenotypes were derived over 2 years of social media posts prior to intake. These AI-based predictions were compared to predictions based on a standard structured interview scale taken at intake, the Addiction Severity Index 6th edition [22] (ASI). We hypothesized that the AI-based predictions would outperform the ASI.

The *digital phenotype* was initially used to predict patients' 30, 60, and 90-day treatment outcomes into three categories: whether they: (a) remained abstinent, (b) dropped out of treatment before reporting a relapse, or (c) reported a relapse. Figure 1B shows the class distribution across the sample at 30-day increments. Using a random forest model [23] on top of the deep learning-based digital phenotype, we considered both the ASI and the digital phenotypes as predictors on top of demographics (age, gender, and race) within 10-fold cross-validation which avoids overfitting by evaluating model accuracy on samples not used during model training [24]. Accuracy was represented as the average for all three categories of the Area Under the Receiver Operating Curve [25] (ROC-AUC), a characterization of the false-positive to true-positive rate (50% indicates chance and 100% indicates perfect prediction). Full details of the techniques employed follow.

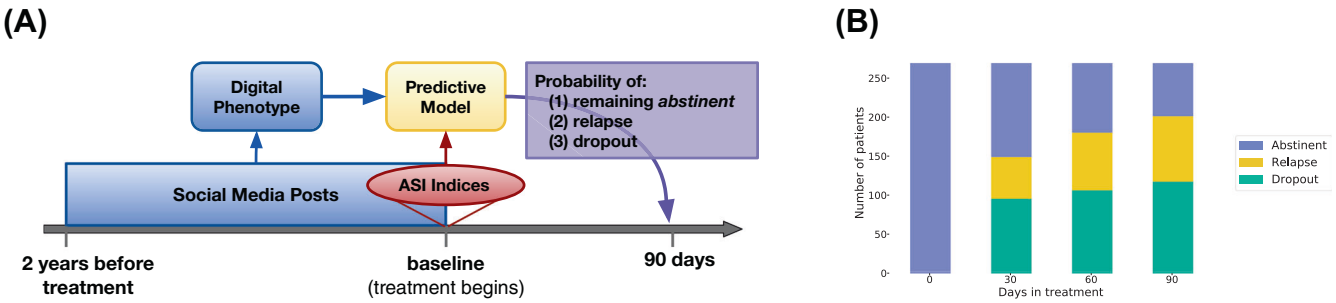

Fig. 1 Study design and descriptive information. A Depiction of study analytic setup: language use patterns from 2 years prior to the start of treatment (baseline) are used within a survival analysis-based machine learning model to forecast the probability of relapse at a given week for each individual. Addiction Severity Index (ASI). **B** Cohort classes by the number of days in treatment.

Table 1. Demographics in the study.

Demographics	Total sample ($N = 504$)	Sample with 200-word restrictions			
		Total ($N = 269$)	Remained abstinent ($N = 68$)	Relapsed ($N = 83$)	Dropped out ($N = 118$)
Age (mean, SD)	33.1 (9.7)	33.2 (9.4)	37.1 (9.8)	32.1 (8.9)	31.7 (8.7)
Sex, % Male	69.3%	63.6%	60.3%	62.7%	66.1%
Race, % Black	59.7%	62.8%	66.2%	72.3%	54.2%
Ethnicity, % Hispanic	13.5%				
Never married	81.2%				
Social media language					
Number of words (mean, SD)	43,395 (7709)	4488 (8850)	4171 (7879)	2863 (4423)	5814 (11,178)
Drug use history					
Drug treatment attempts (mean, SD)	3.8 (4.1)				
Reason entering treatment					
Alcohol	9.9%				
Marijuana	28.0%				
Sedatives	2.2%				
Cocaine/Crack	18.8%				
Stimulants	1.0%				
Hallucinogens	7.7%				
Heroin	22.2%				
Other Opiates	7.3%				
Other Substances	2.8%				

Participants

We recruited participants from four community drug-free outpatient treatment programs near Philadelphia, Pennsylvania. Two research assistants visited each treatment facility and approached patients for participation. Patients were eligible to join the study only if they met the following criteria: above 18-year-old, on their treatment intake day or within the first week of treatment entry, U.S. residents, social-media users, and having no cognitive impairments. Besides these, participants were excluded if their enrollments were mandated by the judicial system or did not have a Facebook account (see details in ref. [26]). Institutional review board (IRB) approval for this study was obtained from the University of Pennsylvania and informed consent was obtained from all participants. Consenting participants ($N = 504$, see Table 1) completed a 30- to 45-minute baseline assessment which included answering a structured interview (ASI; see below) as well as sharing Facebook data (through an automated download using the official Facebook Application Programming Interface). Participants remained in the study for a maximum of 26 weeks (i.e., 6 months) post admission. Each week participants took a short online survey, which asked about relapsing as well as alcohol and drug consumption since the previous survey. Participants were compensated up to $134 for their participation in the study ($30 for completing baseline assessment, $4 for completing each weekly substance use report; see Supplemental Materials for full question text for weekly substance use reports). Participants were informed that their payment was not dependent on how they responded to the weekly substance use reports. We further exclude participants who have less than 200 words in their 2-year Facebook status updates (including links) before they enter the study, which lead to a sample of 269 being included in our data analysis. Those who met this criterion did not differ significantly in demographics or in drug history severity, while the overall sample was majority male and African American (Table 1).

Addiction Severity Index and treatment outcome

Participants were administered the Abridged Addiction Severity Index 6th edition (ASI-6) by a trained research assistant. The Abridged ASI-6 was used to produce 3 recency scores for each participant: psychiatric (ASI-psych); alcohol (ASI-alcohol); and drugs (ASI-drug). The ASI-alcohol and ASI-drug scores are composed of 45 items related to recent alcohol and other drug use, problems, and service utilization. The ASI-psych scores are composed of 21 items related to a variety of recent specific psychiatric symptoms, associated distress, impairment, and service utilization. ASI-6 RSSs were calculated following the author's instructions [22]. See Supplemental Materials for more details on the ASI.

3-category outcome. We defined an outcome variable containing three possible values using the self-report survey data: *remained abstinent, relapse*, and *dropout*. All three outcomes were defined across 30, 60, and 90-day time periods, relative to a participant's enrollment date. *Remained abstinent* was used if the participant reported remaining in treatment and did not report a relapse for the length of the given time period. *Relapse* was used for participants answering *yes* to the question "Did you relapse in the past week?" at any point within the 30, 60, or 90-day period. Finally, *Dropout* was used for all other participants – those whose last date of participation was prior to the given time period and who did not report a relapse within the time period. For example, if a participant never reported relapse and answered their last survey at 45 days past enrollment, then *dropout* at 30 days was false (i.e., they either *remained abstinent* or *relapsed* depending on if they reported a relapse), but *dropout* at 60 and 90 days was true.

4- and 2-category outcomes. We also defined an alternative 4-category outcome variable by splitting the relapse variable into two categories: *relapse-in* and *relapse-out*. *Relapse-in* consists of those that reported relapse and then continued participation. *Relapse-out* was thus used for those that reported a relapse but then otherwise met the criteria for *dropout*. Altogether this results in 4 possible variables for each of the 30, 60, and 90-day time periods: *remained abstinent, relapse-in, relapse-out*, and *dropout*. Finally, our 2-category outcome consisted of simply grouping the *remained abstinent* and *relapse-in* groups into a single *stay-in treatment* category while the *relapse-out and dropout* outcomes were combined into a single *dropped-out* category. See Fig. 3C for a depiction of how the 3-, 4-, and 2-category outcome variables are related to each other and sample sizes within each category.

Facebook language as a digital phenotype

Facebook language data was collected from two sources: status posts and link posts. Status posts, often called *wall posts* or *status updates*, are longer pieces of text users post on their Facebook timeline. Links posts are links shared by participants that include free text along with the original link–this free-form text is often shorter than a typical status post. Following previous work using Facebook language to predict records of depression [13], Facebook language data was limited to the two years prior to entering treatment, and participants must have posted at least 200 words across both statuses and links.

To extract a digital phenotype for each participant, we used a modern deep learning technique, the transformer language model. Such and have recently achieved state-of-the-art performance in language-based assessment of personality [17] and degree of depression [21]. Specifically, we used the Bidirectional Encoder Representation from Transformer (BERT-large) model following the approach of Matero et al. [21] and Ganesan et al. [17], and implemented within the Differential Language Analysis Toolkit (DLATK [27]). We first note that, a full post history (or even single post) on Facebook may be longer than the maximum token limit for the BERT large model (i.e., a maximum of 512 tokens). Thus, we begin by first splitting each user's posts into a sequence of sentences. Then, for each sentence, the pretrained BERT model (i.e., the BERT large model, uncased) is run which produces an embedding for each word based on its context (i.e., the surrounding sequence of words)–a vector representation with 4096 dimensions. The 4096 dimensional vector representation is taken from the last four hidden layers of the BERT model (1024 ×4), which tends to be more task specific, whereas earlier hidden layers contain more information about sentence structure and syntax [28]. DLATK then aggregates the word level BERT embedding across all words within a sentence and then combines the sentence level embeddings using a minimum, maximum, and average of all values. For example, for a user with 100 sentences, the minimum, maximum, and average of each of the 1024 dimensions were recorded, resulting in $1024 \times 4 \times 3 = 12{,}288$ dimensions.

The process of extracting BERT embeddings was applied to both the status updates and post links of the users. Following Matero et al. [21] and Ganesan et al. [17], because our data size was much smaller than the 12,288 dimensions this resulted in, we then applied a dimensionality reduction technique, non-negative matrix factorization, in order to reduce the size of the user representation. The user-level representation built from the status updates was reduced to 35 dimensions, while the post links embeddings were reduced to 15 (given the fact that this type of text is typically shorter than a status update), adding up to 50 total values representing each participant, following the criteria of Ganesan et al. [17] finding ~48–64 feature dimensions ideal for sample size of $N = 269$. Besides BERT, to test the robustness of our approach, we also extract and use other relevant text features (such as RoBERTa [29], BERTweet [30], n-grams), as well as using different classification model (such as support vector machines [31], logistic regression [32]) to test the approach.

We also extracted three measures of linguistic style for both link posts and status updates (for a total of 6 features): average word length, average words per message, and the total number of words used. Due to the limited sample size, we did not explore other linguistic style measures, and, instead, relied on the default measures used in our text analysis package DLATK. Finally, all models utilized demographics as a base feature set in order to control for age, gender, and race disparities in outcomes. These included binary age terciles, binary gender (1 for female, 0 otherwise), and binary race (1 for African American, 0 otherwise). This combination gave us a total of 61 features per participant, which we refer to as the *digital phenotype*.

Statistical analysis

Predictive modeling evaluation. One of the advantages of BERT and similar techniques is that they are pre-loaded to do an excellent job, relative to previous techniques, quantifying language. In other words, they have already done most of the "heavy lifting" in terms of statistically modeling the language. Thus, the final steps of turning our 61-feature *digital phenotype* into predictions of the outcome variable need not use a highly parameterized model. Following Matero et al. [21] we used a random forest model with the extremely randomized trees algorithm [23] to produce a model which estimates probabilities for each treatment outcome given the *digital phenotype*. For models using the ASI baseline, which had only 3 variables, we found ridge penalized logistic regression to be more ideal than the random forest approach (the random forest approach is known to work best with higher dimensional data).

Both the extremely randomized trees and ridge penalized logistic regression approaches employ techniques to avoid overfitting when training the statistical models. However, a cross-validation technique is necessary to establish the accuracy of such predictive models [24, 33] whereby test data is held out (not input) during the model fit (training), and then the model is run over the held-out data to test for accuracy. We used a 10-fold cross-validation technique to evaluate the different combinations of the digital phenotype and ASI scores. We divided the sample of 269 participants into 10 random chunks, or folds, such that each fold had roughly the same distribution of classes (stratified random folds). This "n-Fold" cross-validation technique was preferred to leave-one-out cross-validation [33] due to the computational demands of training 238 models versus only 10 models for n-fold and previous work suggest n-fold cross-validation is less prone to overfit [24]. For each of the 10 folds, we trained a model on the other 9 folds, and then applied the trained model to the remaining fold, in such a way that each fold was held out exactly once. Thus, the test fold was only observed once in order to establish accuracy during cross-validation. Hyperparameters (number of estimators in the case of random forests and regularization penalty for the SVM and linear models used in sensitivity testing) were set during each fold based on tests over the training set.

Dropout risk. We evaluated the digital phenotype predictions as a dropout risk score. Simulating clinical application at baseline, in addition to the digital phenotype, we included ASI scores and demographic variables which are also available at treatment intake. The random forest model outputs a probability of drop-out per subject (the same probabilities used for the AUC as described above). We divided these into four uniformly-sized "risk quartiles": lowest risk quartile: [0%, 38%], 2nd risk quartile: [39%, 55%], 3rd risk quartile: [56%, 68%], and highest risk quartile: [69%, 95%]. We utilized the same 10-fold cross-validation technique previously mentioned such that individuals' risk quartile was always predicted from a random forest model that had not seen the individual in training. Thus, we labeled each participant with a risk score from the predictive model without any knowledge of the patient's final outcome – and only using their language use on social media in the 2 years prior to baseline. We then calculated the proportion of each of the 4 quartiles that remained in treatment at 30, 60, and 90-days post-intake and placed such proportion in a survival plot.

RESULTS

Predictive model evaluation

Predictions of 90-day SUD treatment outcomes from the digital phenotype were greater than those from ASI (see Fig. 2: AUC of 0.725 versus AUC of 0.658; single-tailed permutation test $p < 0.001$). That is, relapse and drop-out from SUD treatment could be predicted substantially better by analyzing social media posts using the AI-based method than from a standard psychometric assessment tool administered at intake, supporting our hypothesis. Accuracy scores were the largest when combining the ASI and digital phenotype, and significantly greater than ASI alone (AUC of 0.739; $p < 0.001$). This data suggests knowing the digital phenotype at treatment intake can aid, beyond simply knowing the ASI, in understanding who is likely to relapse or drop out of treatment during the first 90 days.

We can also examine the linguistic features that best predicted these outcomes. Figure 3A depicts the three outcome categories according to the standardized mean values of their digital phenotypes. Remained abstinent and dropout categories appear to have distinct signatures. Relapse, though, seems to sometimes mirror abstinence, and other times it does not. When comparing the Euclidean distance between each signature, depicted on the right side of the plot, we see that abstinent and dropout are furthest apart, while relapse is somewhere in between the two. This data is corroborated by the observation depicted in Fig. 3B that relapse (AUC = 0.659) was harder to predict than abstinence (AUC = 0.724) or dropout (AUC = 0.752).

While dropout is also often indicative of relapse [34], the outcomes showed both "relapse" and "abstinent" groups continued to engage in treatment. To be in the relapse group, one must have reported it, while the dropout group has severed all communication with the treatment providers. Therefore, it is possible that the relapse group is more like the abstinent group than the non-compliant group.

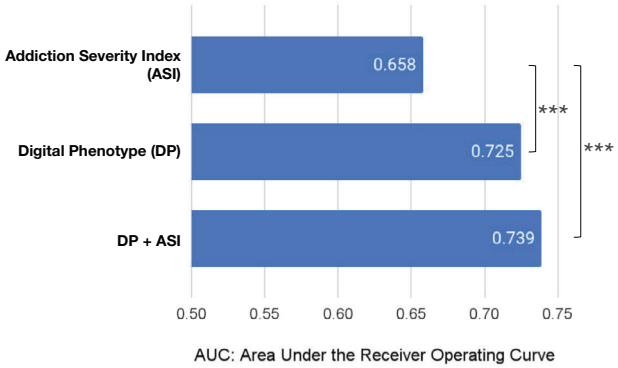

Fig. 2 Overall accuracy predicting treatment outcome at 90 days. We categorized patients into categories of *abstinent*, *relapse*, and *dropout* using only information available at baseline: the digital phenotype (DP) and the addiction severity index (ASI; the clinical measure), each alone, as well as combined (DP + ASI). ASI scores were from treatment intake, language was from two years before treatment until intake. AUC scores were evaluated out-of-sample using a 10-fold cross-validation of predictions from a random forest model. Difference between ASI and DP and ASI and DP + ASI was significant (single-tailed permutation test, $p < 0.001$). Difference between DP and DP + ASI was not significant. See Supplementary Table S1 for sensitivity and specificity.

In the next analysis, we investigated an alternative conceptualization of the relapse outcome inspired by work demonstrating that relapse, itself, is not a remarkable event for SUD recovery [35]. We categorized those in the relapse outcome group based on whether they remained in treatment after the relapse (*relapse-in*) or dropped out before 90 days (*relapse-out*). This categorization allows those who relapsed and stayed in treatment to be further grouped with those remaining abstinent into a single "remained in treatment" outcome, while those who relapsed and then dropped out are grouped with those who dropped out. Figure 3C shows the sample after it has been divided into four outcomes (*abstinent, relapse-in, relapse-out*, and *dropout*) and then further grouped into two outcomes (*remained in treatment, dropped out)*. Given the additional outcome, one would expect, simply due to chance, lower accuracy––*unless* this demarcation is more appropriate (i.e., the phenotypes for those who end up in the *relapse-in* category are distinguishable from those who end up categorized as *relapse-out*). We then compared predictive accuracies of the digital phenotype (in terms of AUC) for these categorizations. Results are depicted in Fig. 3C. Predicting 4 outcomes (AUC = 0.792), with relapse divided into two, had significantly greater AUC than 3 outcomes (AUC = 0.739; $p < 0.01$). Further, simply predicting 2 outcomes, *remained in* or *dropped out*, had the highest levels of prediction (AUC = 0.806). In the robustness test, in which we tested the method with different text features (e.g., RoBERTa, BERTweet, n-grams) and classifying models (e.g., SVM, logistic regression) and reported the results in Supplementary Fig. S1. The results show that the predictive capability of combined ASI and digital phenotypes features hold for other text features and learning models.

Dropout risk: simulating clinical risk score at intake

We next evaluated the digital phenotype predictions as a dropout risk score. Using predicted outcome probabilities as risk scores simulates use in a clinical application at baseline [33]. The full model using the digital phenotype, ASI, and demographics were able to predict dropout with AUC = 0.81 (sensitivity: 0.81; specificity: 0.67) at 90 days. Figure 4A plots the full ROC-AUC curve for the final dropout probability estimates, with the x-axis representing the false-positive rate, the y-axis representing the

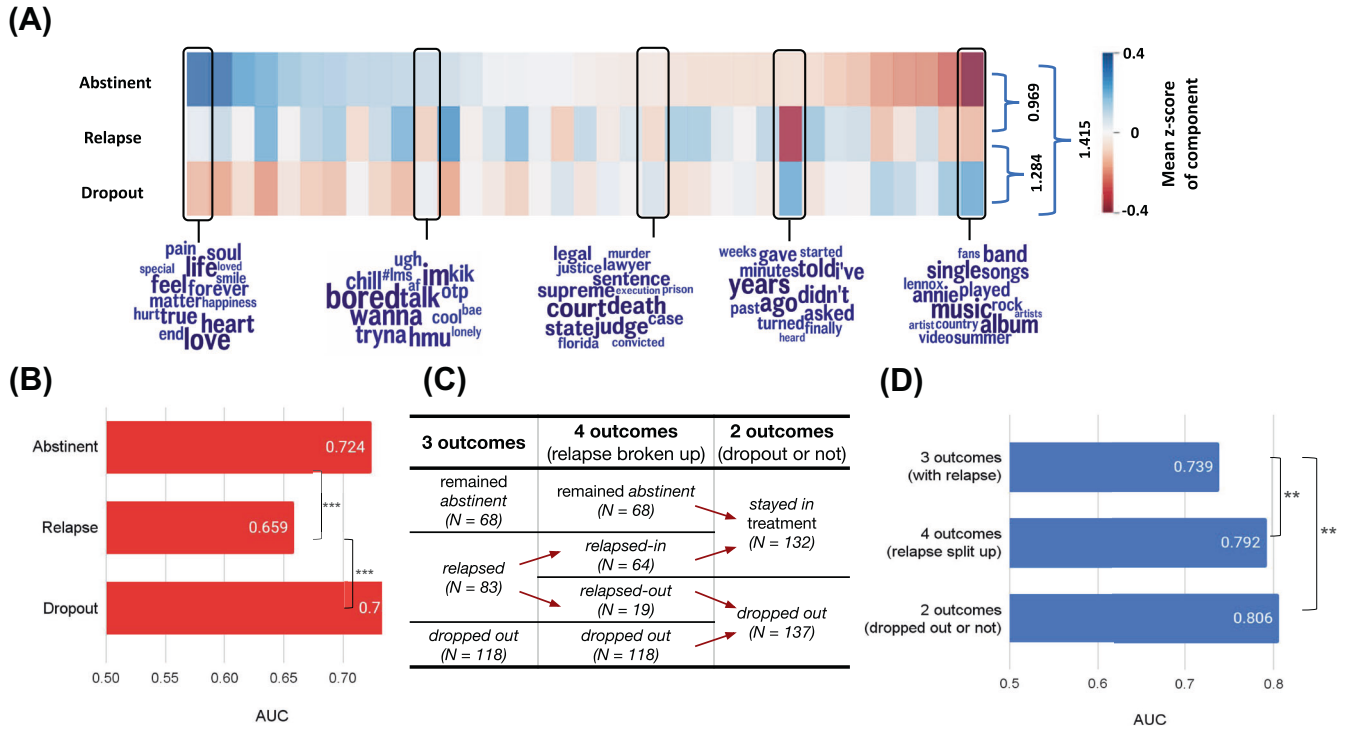

Fig. 3 Prediction of future abstinence, relapse, and dropout at 90 days. A Expression of digital phenotypes. Deep language component scores across relapse, non-compliance, and abstinence by participants from the given class. Each column corresponds to a component and colors correspond to the mean z-score (standardized score) for the class. Components were sorted by their score for the abstinent class. **B** Prediction accuracies for each of the three treatment outcomes. **C** Depiction of division of relapse into subcategories: *relapse-in* and *relapse-out*. **D** Prediction accuracy when using four outcomes (*abstinent*, *relapsed and stayed in treatment*, *relapsed and then dropped out*, or *dropped out*) and two outcomes (*remained in* or *dropped out*) where relapse is broken up as depicted in panel **C**. Accuracy is significantly greater for both the 4-class and 2-class divisions of outcomes than for the 3-class division. ***$p < 0.001$, **$p < 0.01$.

true-positive rate, and the diagonal representing random chance. While slightly favoring the detection of those remaining in treatment, the curve was fairly uniform in terms of false-positive to true-positive rates. Thus, there is strong separability regarding who will be in treatment at 90 days.

We can directly examine the pre-treatment risk separation by considering the proportion of our participants who remained in treatment, broken into the four quartiles of strata based on pre-baseline digital phenotype and demographics in Fig. 4B. By 30 days, a large separation between risk quartiles is established, and, on average, the quartiles continue to separate going into 60- and 90-day outcomes (difference between highest and lowest quartile, $p < 0.001$). We also see that approximately 50% of the two highest risk quartiles dropout of the study by 30 days. Thus, we see that this sample is similar to previous studies in that dropout is the norm rather than the exception [1–3].

Further validation of the risk score can be examined based on its ability to distinguish the 4 treatment outcomes at 30-, 60-, and 90- days. In Fig. 4C, nearly all individuals who ended up within the *relapse-in* category were identified as low risk, using only data pre-treatment, while most individuals ending up in the relapse-out category by 90 days had been identified as high risk. This analysis simulated risk scores that could have been made available for patients while entering a particular treatment center and, for those in the top or bottom quartile of risk, their eventual treatment status could have been known at intake with only moderate uncertainty.

DISCUSSION

The digital phenotype extracted from language on Facebook combined with standard intake data, showed strong validity as a risk assessment tool for SUD treatment dropout. We were able to show such data can capture many distinctions between treatment outcomes. In fact, while we were able to predict all of the three classes better with the digital phenotype than the ASI, relapse was significantly harder to predict and we found that dividing it into those that relapsed but remained in treatment versus those that relapsed and dropped out, resulted in greater accuracy, suggesting a potential lack of utility of focusing on relapse as a whole. With most adults in the US using social media regularly (70%) [36] and little variation across demographics [7] or SUD diagnosis [8, 9], such techniques could transform intake assessment and clinical practice when applicable (i.e., patients who engage with these platforms as opposed to those who anonymously scroll). Specifically, digital phenotypes extracted from social media language can be a screening tool for identifying SUD patients at high risk for drop-out at the beginning of treatment engagement. This work also sets a foundation for dynamic continuous monitoring during treatment to better understand the day to day factors that go beyond whether someone is likely to succeed on day one.

While the digital phenotype from social media offers a new tool for intake risk assessment, it is important to keep in mind its limitations. First, the *abstinent* and *relapse* outcomes relied on patient self-report. Such results should be replicated with biological measures of drug use, as these measures cannot be influenced by social expectations or recall. Another limitation is the dynamic nature of language across time, space, and cultures. The model built here should work reliably for other members of the same, majority middle-aged and African American, population. One would need to develop and validate a similar model for populations from different cultures and over time. Similarly, this study is also limited by its sample size – only 269

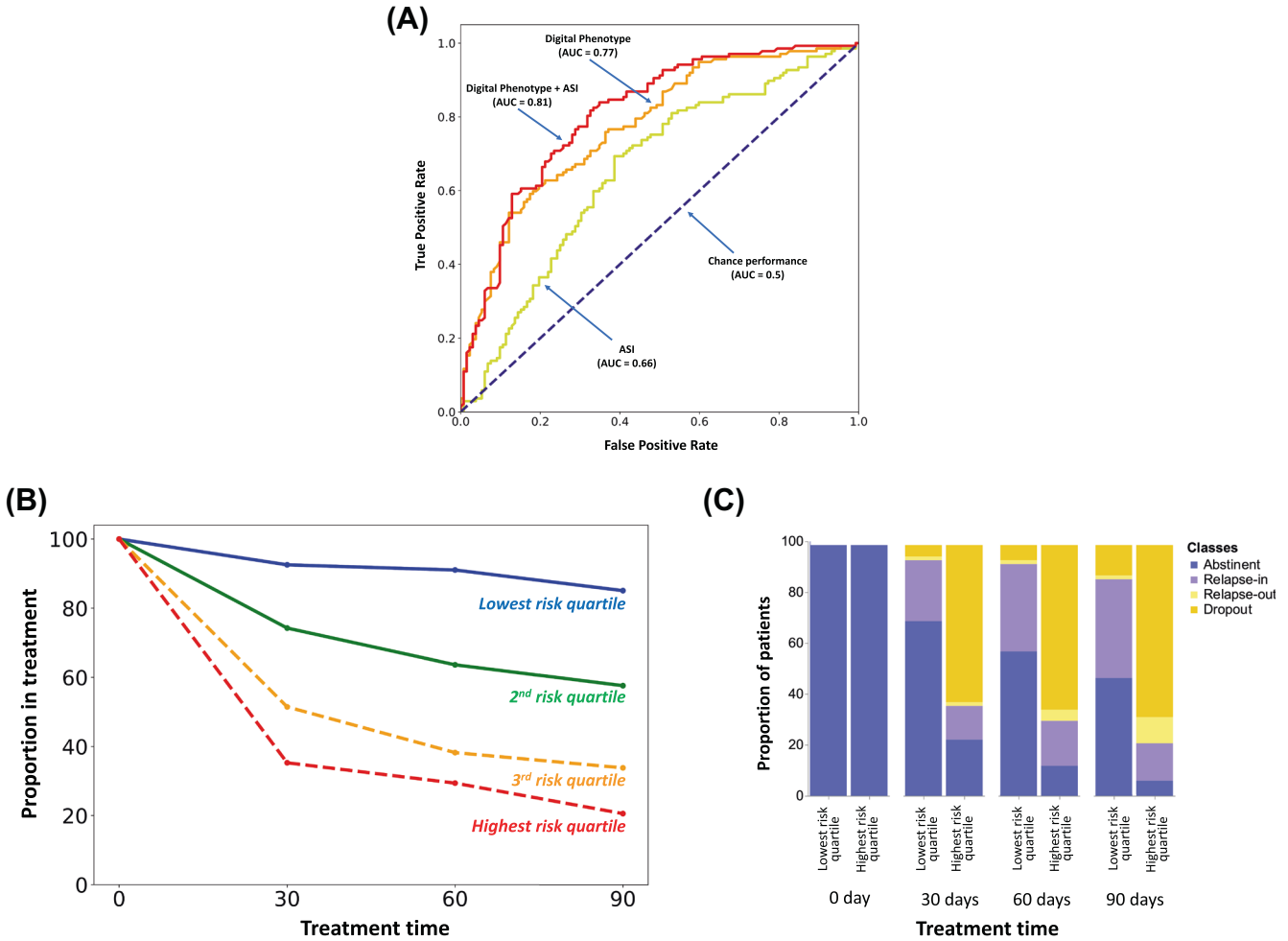

Fig. 4 Performance of our full risk assessment model which predicts the probability of dropout by 90 days. A Receiver Operating Characteristic (ROC) Curve for model predictions of Dropout at 90 days; **B** Proportion in treatment by risk quartile; **C** Proportion of the four outcomes within the lowest risk and highest risk quartiles at 0, 30, 60, and 90 days of treatment.

participants met the required 200 word minimum across their Facebook data. While similar sample sizes and word minimums have been used in other social media-based tasks [10, 17], further validation is needed to test whether the results generalize across different samples. The digital phenotype using BERT, only offers a limited view for interpreting the most predictive linguistic features. The future analysis would need to use models of specific psychological constructs to better understand the latent variables that the model is tracking. Finally, such risk assessments are only available to those with enough social media data (just over half of the participants in our sample).

Digital phenotyping is already being leveraged *en masse* by corporations and some governments, but its use has been limited in medical environments. This study demonstrates the feasibility of predicting treatment outcomes, by analyzing the language behavior of SUD patients on social media, and outperforming psychometric interview-based scales, which are the current standard. This AI-based method of digital phenotyping has significant clinical implications. First, AI-based linguistic analysis may suggest new variables of importance for SUD treatment outcomes. Selective linguistic features and language use patterns may be identified as strong signals of certain behaviors and treatment outcomes, tracking these markers during clinical encounters and counseling sections could lead to quicker interventions. Second, AI-based prognostic assessment and intervention tools could be developed to automatically screen and intervene with the patients in real time, which would increase the efficiency and accessibility of SUD assessments and treatments and reduce the cost of healthcare in this process.

REFERENCES

1. McLellan AT, Lewis DC, O'Brien CP, Kleber HD. Drug dependence, a chronic medical illness: implications for treatment, insurance, and outcomes evaluation. JAMA J Am Med Assoc. 2000;284:1689–95.
2. Paliwal P, Hyman SM, Sinha R. Craving predicts time to cocaine relapse: further validation of the Now and Brief versions of the cocaine craving questionnaire. Drug Alcohol Depend. 2007;93:252–59.
3. Brecht M-L, Herbeck D. Time to relapse following treatment for methamphetamine use: a long-term perspective on patterns and predictors. Drug Alcohol Depend. 2014;139:18–25.
4. Volkow ND. Personalizing the treatment of substance use disorders. Am J Psychiatry. 2020;177:113–16.
5. Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA. 2017;318:1215–16.
6. Kwako LE, Bickel WK, Goldman D. Addiction biomarkers: dimensional approaches to understanding addiction. Trends Mol Med. 2018;24:121–28.
7. Ashford RD, Lynch K, Curtis B. Technology and social media use among patients enrolled in outpatient addiction treatment programs: cross-sectional survey study. J Med Internet Res. 2018;20:e84.
8. Curtis BL, Ashford RD, Magnuson KI, Ryan-Pettes SR. Comparison of smartphone ownership, social media use, and willingness to use digital interventions between generation z and millennials in the treatment of substance use: Cross-sectional questionnaire study. J Med Internet Res. 2019;21:e13050.

9. Bergman BG, Greene MC, Hoeppner BB, Kelly JF. Expanding the reach of alcohol and other drug services: Prevalence and correlates of US adult engagement with online technology to address substance problems. Addict Behav. 2018;87:74–81.
10. Kern ML, Park G, Eichstaedt JC, Schwartz HA, Sap M, Smith LK, et al. Gaining insights from social media language: methodologies and challenges. Psychological Methods. 2016;21:507–25.
11. Son Y, Clouston SAP, Kotov R, Eichstaedt JC, Bromet EJ, Luft BJ, et al. World Trade Center responders in their own words: predicting PTSD symptom trajectories with AI-based language analyses of interviews. Psychological Med. 2021;53:1–9.
12. Coppersmith G. Digital life data in the clinical whitespace. Curr Directions Psychological Sci. 2022;31:34–40.
13. Eichstaedt JC, Smith RJ, Merchant RM, Ungar LH, Crutchley P, Preoţiuc-Pietro D, et al. Facebook language predicts depression in medical records. Proc Natl Acad Sci. 2018;115:11203–08.
14. Youyou W, Kosinski M, Stillwell D. Computer-based personality judgments are more accurate than those made by humans. Proc Natl Acad Sci. 2015;112:1036–40.
15. Park G, Schwartz HA, Eichstaedt JC, Kern ML, Kosinski M, Stillwell DJ, et al. Automatic personality assessment through social media language. J Personal Soc Psychol. 2015;108:934.
16. Curtis B, Giorgi S, Buffone AE, Ungar LH, Ashford RD, Hemmons J, et al. Can Twitter be used to predict county excessive alcohol consumption rates? PloS One. 2018;13:e0194290.
17. Ganesan AV, Matero M, Ravula AR, Vu H, Schwartz HA. Empirical evaluation of pre-trained transformers for human-level NLP: the role of sample size and dimensionality. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics; 2021. p. 4515–32. https://doi.org/10.18653/v1/2021.naacl-main.357.
18. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2019, p. 4171–86.
19. Boyd RL, Schwartz HA. Natural language analysis and the psychology of verbal behavior: the past, present, and future states of the field. J Lang Soc Psychol. 2021;40:21–41.
20. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. arXiv. 2021. https://arxiv.org/abs/2108.07258.
21. Matero M, Idnani A, Son Y, Giorgi S, Vu H, Zamani M, et al. Suicide risk assessment with multi-level dual-context language and BERT. In Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology. Minneapolis, Minnesota: Association for Computational Linguistics; 2019, p. 39–44.
22. Cacciola JS, Alterman AI, Habing B, McLellan AT. Recent status scores for version 6 of the Addiction Severity Index (ASI-6). Addiction 2011;106:1588–602.
23. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn. 2006;63:3–42.
24. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning data mining, inference, and prediction, 2nd ed. Springer Series in Statistics. New York, NY: Springer New York; 2009. p. 219–59.
25. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27:861–74.
26. Liu T, Giorgi S, Yadeta K, Schwartz HA, Ungar LH, Curtis B. Linguistic predictors from Facebook postings of substance use disorder treatment retention versus discontinuation. Am J Drug Alcohol Abuse. 2022:48(5):573–85.
27. Schwartz HA, Giorgi S, Sap M, Crutchley P, Ungar L, Eichstaedt J. Dlatk: Differential language analysis toolkit. In Proceedings of the 2017 conference on empirical methods in natural language processing: System demonstrations; 2017, p. 55–60.
28. Rogers Anna, Kovaleva Olga, Rumshisky Anna. "A primer in BERTology: what we know about how BERT works." Transactions of the Association for. Computational Linguist. 2021;8:842–66.
29. Yinhan L, Myle O, Naman G, Jingfei D, Mandar J, Danqi C, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. 2019. https://arxiv.org/abs/1907.11692.
30. Nguyen DQ, Vu T, Nguyen AT. BERTweet: a pre-trained language model for English Tweets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Association for Computational Linguistics; 2020, p. 9–14.
31. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97.
32. McCullagh P, Nelder JA. Generalized linear models. Monographs on Statistics and Applied Probability, 2 edn. Vol. 37. Chapman and Hall: CRC press; 1989.
33. Rammos A, Gonzalez LAN, Weinberger DR, Mitchell KJ, Nicodemus KK. The role of polygenic risk score gene-set analysis in the context of the omnigenic model of schizophrenia. Neuropsychopharmacology. 2019;44:1562–69.
34. Ashford RD, Giorgi S, Mann B, Pesce C, Sherritt L, Ungar L, et al. Digital recovery networks: characterizing user participation, engagement, and outcomes of a novel recovery social network smartphone application. J Subst Abus Treat. 2020;109:50–55.
35. Perrin A, Anderson M. Share of US adults using social media, including Facebook, is mostly unchange since 2018. 2019. https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/.
36. Social media fact sheet. 2021. https://www.pewresearch.org/internet/fact-sheet/social-media/.

ACKNOWLEDGEMENTS

The corresponding author, Dr. Brenda Curtis, had full access to all the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis.

AUTHOR CONTRIBUTIONS

Conceptualization: BC, SG, LU, HAS. Methodology: BC, SG, LU, HAS. Investigation: BC, SG. Visualization: SG, HV, HAS. Funding acquisition: BC, LU, HAS. Project administration: BC, LU. Supervision: BC, LU. Writing – original draft: BC, SG, HAS. Writing – review & editing: BC, SG, LU, DY, TL, KY, HAS. Software: SG, HV, HAS. Validation: TL, HV. Formal Analysis: SG, TL, HAS. Resources: BC. Data Curation: LU, HAS.

FUNDING

This research was supported by: Intramural Research Program of the NIH, NIDA (BC), Templeton Research Trust (LU, HAS), National Institute on Drug Abuse Grant # R01DA039457 (BC, LU), National Institute on Alcohol Abuse and Alcoholism # R01 AA028032/AA (HAS, LU). The authors declare that they have no competing interests. This study was funded by the Intramural Research Program of the National Institutes of Health (NIH), National Institute on Drug Abuse (NIDA).

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41386-023-01585-5.

Correspondence and requests for materials should be addressed to Brenda Curtis.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.