Supplemental Material for: AI-based Analysis of Social Media Language Predicts Addiction Treatment Dropout at 90 Days

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

Affiliations:

[1]Intramural Research Program, National Institute on Drug Abuse, Baltimore, USA.

[2]Positive Psychology Center, University of Pennsylvania, Philadelphia, USA.

[3]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, USA.

[4]Department of Computer Science, Stony Brook University, USA.

[5]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, USA.

*Corresponding author. Email: brenda.curtis@nih.gov.

Methods

Participants

We recruited participants from four community drug-free outpatient treatment programs. Two research assistants visited each treatment facility and approached patients for participation. Patients were eligible to join the study only if they were recently admitted: on their treatment intake day or within the first week of treatment entry. Consenting participants completed a 30- to 45-minute baseline assessment, which included sharing Facebook data (through an automated download using the official Facebook Application Programming Interface) and answering the Abridged Addiction Severity Index 6th edition (ASI). Both sharing Facebook data and the ASI were mandatory for study participation. Participants also completed a TimeLine Follow-Back (TLFB), reporting their alcohol and drug use for the two weeks prior to entering treatment. Institutional review board (IRB) approval for this study was obtained from University of Pennsylvania and informed consent was obtained from all participants.

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. In particular, participants were asked to respond to the questions: “Did you relapse in the past week?”, “Did you drink alcohol or use any other drugs during the past week?”, “Did you have more than *N* drinks at any one time in the past week?”, and “Please select all of the drugs you used in the past week? (Select all that you used)”. Here *N* is 3 for women and 4 for men.

Addiction Severity Index and Treatment Outcomes

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 [1]. The ASI, including the psychiatric, alcohol, and drugs recency scores, have been found to be reliable and valid [2-3] across several diverse populations, including treatment centers, psychiatric patients, and those who are unhoused [2, 4-5].

We define two binary treatment outcomes using the self-report survey data: *relapse* and *dropout*. Both outcomes are defined across 30, 60, and 90-day time periods, relative to a participant's enrollment date. *Relapse* is set to 1 (0, respectively) for participants answering *yes* (*no*, respectively) to the question “Did you relapse in the past week?” at any point within the 30, 60, or 90-day period. If a participant did not respond to a survey in a given week, then *relapse* is set to null. *Dropout* is defined by looking at the date of the participants’ last survey. If the participant answered a survey at any point later than the given time period, then *dropout* is set to 1 (0, otherwise). For example, if a participant answered their last survey at 45 days past enrollment, then *dropout* at 30 days is 0, *dropout* at 60 days is 1, and *dropout* at 90 days is 1. In order to make these mutually exclusive classes, if a participant reports a relapse within a given time period, then dropout is set to 0. Thus, the final definition of *dropout* is a participant who did not relapse and also did not answer a survey outside of the specific time period.

Predictive Modeling

Sensitivity and specificity are shown in Table S1 for the three class prediction results (i.e., Figure 2 in the manuscript).

	30 Days		60 Days		90 Days	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
ASI	0.438	0.718	0.455	0.732	0.434	0.717
DP	0.515	0.762	0.495	0.766	0.501	0.758
DP + ASI	0.530	0.769	0.542	0.779	0.508	0.760

Table S1. Sensitivity and Specificity. Sensitivity and specificity of predicting *abstinent, relapse*, and *dropout* treatment outcomes at 30, 60, and 90-days using only information available at baseline (demographics, the digital phenotype DP, and the addiction severity index ASI). Scores were evaluated out-of-sample using 10-fold cross-validation of predictions from a random forest model.

Robustness test

We tested how the digital phenotype performs with different text features and with different classifying models. Particularly, besides BERT [6] text embeddings, we extracted and measured the performances of Large RoBERTa [7] (average of the second to last layer), Base BERTweet [8] (average of the second to last layer), n-grams (1-to-3 grams). For BERT text features, we also tested them with different learning models, such as SVM and logistic regression. Fig. S1 shows the results. We find that the performances of digital phenotype signals hold across different methods of extracting text embedding features. With other learning models, the performances slightly drop but are still comparable with when using the random forest model.

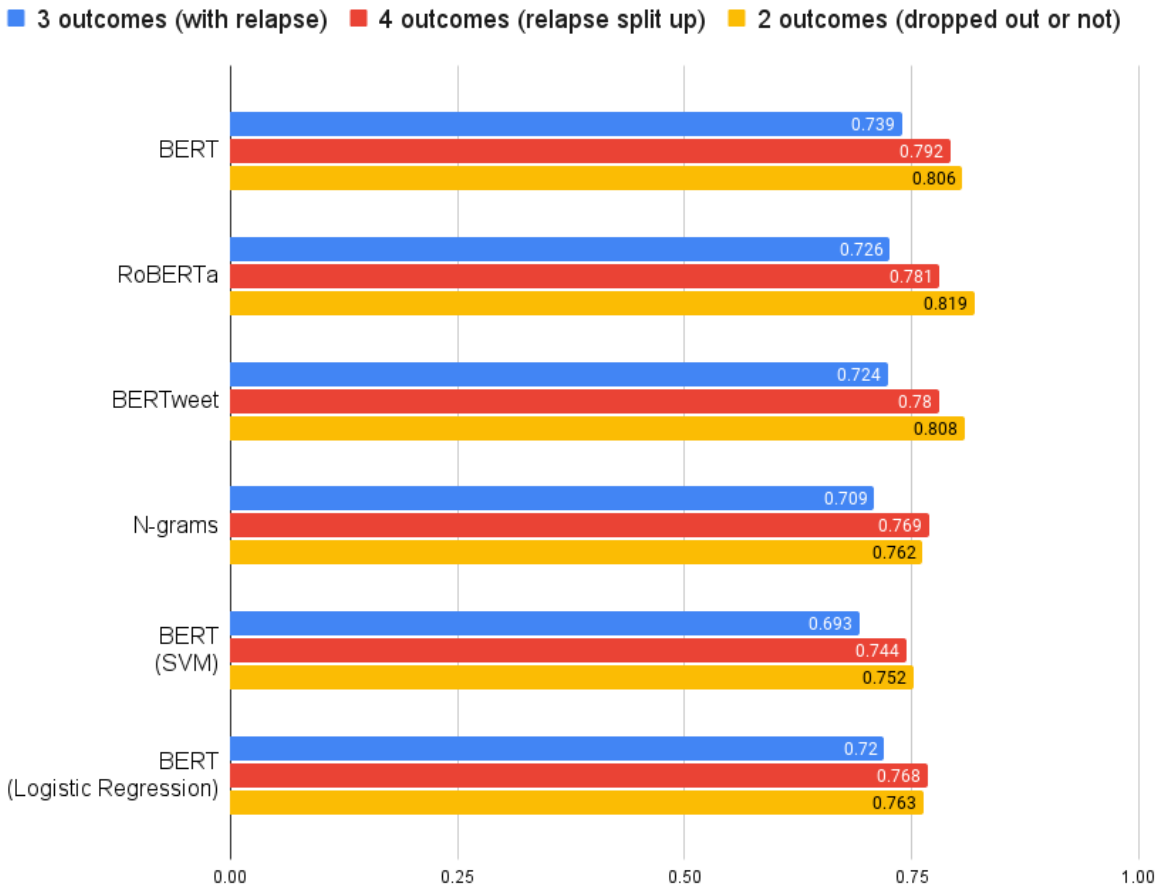

Fig S1. Performances of different digital phenotypes features and learning models. The digital phenotypes text embeddings are extracted and combined with ASI features to predict the class for three outcomes (abstinent, relapsed and dropped out), four outcomes (abstinent, relapsed and stayed in treatment, relapsed and then dropped out, or dropped out) two outcomes (remained in or dropped out). The results are measured with AUC.

Expression of Digital Phenotype

To visualize the linguistic expressions of the digital phenotype dimensions (Fig. 3(**A**) in the main paper), we plotted mean scores for each of the components of the digital phenotype (i.e., the resulting low dimensional representation from non-negative matrix factorization of the deep learning model, BERT-large). The scores were standardized by subtracting the mean and dividing by their standard deviation over our entire sample of participants. Then, means of these standard scores were calculated for each of the three outcome classes: abstinent, relapse, and dropout. Thus, the scores can be interpreted as the Cohen's d of the component for participants in the category versus participants not in the category.

As a descriptive of select digital phenotype (DP) components, we found clusters of words that were most associated with each component. We first built a set of Latent Dirichlet Allocation (LDA) topics [9]. LDA is a generative statistical model used to automatically cluster groups of semantically related words. We used the same set of 55,644 status updates posted in the two years before participants entered treatment. The Java program Mallet [10] was used to build 200 topics. All default settings were used except α, which controls the expected number topics per document (i.e., lower α assumes a smaller number of topics per Facebook statuses update). We set $\alpha = 2$ since our Facebook posts were shorter than typical documents. We also excluded the 40 most frequent words in our data from the LDA process. Finally, we calculated 200 topic loadings (i.e., p(topic|user)) for each of the 269 participants. The topics shown for select DP components are those with the greatest Pearson correlation between the topic score and the component scores. Components were selected to have descriptive topics shown based on having the highest and lowest mean for the abstinent dimension, as well as covering a range of phenotype loadings for the dropout dimension.

Risk Quartiles

We defined the risk score as the estimated probabilities of dropout at 90 days produced from our model using only the DP from prior to intake, as well as intake data (demographics and ASI). These scores were exported from the 10-fold cross validation technique mentioned previously, so they were always produced, out-of-sample, from a model that was trained without observing the participant (in order to simulate application to patients the model had not seen). We then computed quartiles of the risk scores by uniformly dividing the participants into four groups from least to greatest scores. Across all four quartiles, we looked at the proportion of participants remaining in treatment at various times throughout treatment (30, 60, and 90 days). Fig. 4(**B**) in the main text shows the results of this experiment. Finally, we compared the highest risk and lowest risk quartiles and looked at the proportion of participants remaining in treatment in each of our four outcome classes (*abstinent, relapse-in, relapse-out,* or *dropout*). Again, this was done at 30, 60, and 90 days. Results are presented in Fig. 4(**C**) in the main text.

To provide examples of specific language associated with risk scores, Figure S2 displays the words and phrases most associated with a low risk score. Correlation was derived from a Pearson Product-Moment Correlation Coefficient between the term frequency among status updates and model's estimated risk score, overall, from subjects writing at least 200 words in their statuses (N = 206). Terms were restricted to those mentioned by at least 5% of subjects and function words (determiners, conjunctions, pronouns, and prepositions) were excluded, resulting

in consideration of 1,238 terms. All words listed were significantly correlated with risk score (two-tailed $p < 0.05$; Benjamini-Hochberg adjusted for false-discovery rate). For example, mentioning "love", "God", and "family" were associated scores indicating one is likely to remain in treatment. While the goal of this work was to evaluate the utility of the digital footprint for SUD treatment outcome assessment, future work may explore the full extent of specific lexical correlates of treatment outcomes.

Fig. S2. Words and phrases most associated with low risk scores. In this visualization of results from quantitative analysis, size of the term indicates its correlation strength with inverse risk score as derived from the digital phenotype, while color indicates the frequency with which the term is used. All words listed were significantly associated ($p < 0.05$; Benjamini-Hochberg adjusted for false-discovery rate).

References

[1] J. S. Cacciola, A. I. Alterman, B. Habing, A. T. McLellan, Recent status scores for version 6 of the Addiction Severity Index (ASI-6). *Addiction.* **106**, 1588-1602 (2011).

[2] McLellan, A. Thomas, et al. "New data from the Addiction Severity Index reliability and validity in three centers." *The Journal of nervous and mental disease* 173, 412-423 (1985).

[3] Kosten, Thomas R., Bruce J. Rounsaville, and Herbert D. Kleber. "Concurrent validity of the addiction severity index." *The Journal of nervous and mental disease* 171, 606-610 (1983).

[4] Appleby, Lawrence, et al. "Assessing substance use in multiproblem patients: Reliability and validity of the Addiction Severity Index in a mental hospital population." The Journal of Nervous and Mental Disease 185, 159-165 (1997).

[5] Zanis, David A., et al. "Reliability and validity of the Addiction Severity Index with a homeless sample." *Journal of substance abuse treatment* 11, 541-548 (1994).

[6] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (2019).

[7] Liu Yinhan, Ott Myle, Goyal Naman, Du Jingfei, Joshi Mandar, Chen Danqi, Levy Omer, Lewis M., Zettlemoyer Luke, and Stoyanov Veselin. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv abs/1907.11692.

[8] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 9–14, Online. Association for Computational Linguistics.

[9] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993-1022 (2003).

[10] A. K. McCallum, in *http://mallet.cs.umass.edu*. (2002).