Machine learning-assisted treatment selection for smoking cessation

John J. Curtin[aff-1]

[aff-1]Department of Psychology, University of Wisconsin-Madison

Author note

Correspondence concerning this article should be addressed to .

Abstract

This study found some pretty cool results that have both high impact and important clinical implications. For example …

Keywords: Substance use disordersPrecision mental healthCigarette smokingMachine learningTreatment selection

Introduction

Precision mental health

Precision mental health is the application of the precision medicine paradigm to mental health conditions. Precision medicine and precision mental health aim to address an important problem in traditional treatment selection: what works best at a population level does not necessarily work best for a given patient. For example, although treatment A may be more effective than treatment B across the population, it may be that treatment B is markedly more effective for a specific patient.

Rather than relying on population-level efficacy, precision mental health seeks to guide treatment selection using individual difference characteristics that are likely to predict treatment success for each patient.2 Successful precision mental health would increase the likelihood of treatment success for each patient because each patient receives the treatment predicted to work best for them. It would also improve treatment effectiveness rates across the population because each treatment is administered only to the patients for whom that treatment is expected to be their best option.

In addition to offering improved effectiveness, precision mental health approaches may be more resource-efficient. Clinical trials to develop and validate new treatments are expensive, resource-intensive, and extremely slow (CITE?). These costs may also produce a treatment that is no better than existing treatments, or potentially ineffective altogether. In contrast, by seeking to

optimize existing treatments and direct them to the *right* patients, precision mental health stands as a cost-effective alternative poised for more immediate impact to patients.

Researchers have pursued precision mental health – and precision medicine broadly – for decades. In medicine, emphasis on personalizing treatments has grown rapidly with the ascendancy of advanced genetic methods such as genome-wide association studies and polygenic scores3,4. Within precision mental health, an early example comes from the substance use disorder (SUD) domain: the Project MATCH Research Group attempted to match people with alcohol use disorder to a particular treatment based on individual differences such as gender, social support, or symptom severity5,6. Many researchers have followed in their footsteps as the understanding has grown that neither mental health diagnoses nor treatments are one-size-fits-all7. Efforts thus far have often focused on tailoring treatments at the group level; in other words, identifying a (single) factor that divides individuals within a diagnostic category into subgroups that can be treated differently7.

Despite these opportunities for advances, however, precision mental health research has progressed with limited success. Extant research has not yet enabled reliable recommendations for treatment selection even at the group level - let alone for an individual patient. These patient-level predictions are required for clinical implementation; our goal in clinical science is to predict behavior such that we can apply findings to a new patient.

One reason for this slow progress is that many factors influence a complex clinical phenomenon like treatment success. Thus, any single feature (i.e., predictor variable) cannot account for more than a small portion of the variance in treatment success. This idea is perhaps

comparable to the shift in understanding within genetics: research has moved away from candidate gene studies to polygenic approaches that rely on small contributions from many genes (Bogdan et al., 2018). Unfortunately, traditional analytic techniques have often limited the ability to consider more than one or a few features simultaneously. These limitations have also prevented considering concurrently features across constructs (e.g., demographics, psychological traits, environmental variables). Therefore, models have failed to capture the real-world complexity underlying these clinical phenomena.

Moreover, because researchers using traditional analytic techniques typically develop and evaluate their precision mental health models in a single sample, the models may become very overfit to that sample. Consequently, they do not generalize well to new patients that were not used for model development. This problem is particularly concerning because clinical implementation of precision mental health requires that these models provide accurate recommendations about treatment selection for *new* patients rather than explaining treatment success within the study sample.

These pitfalls interact with each other. To capture sufficient complexity to predict treatment success, we need to increase the total number of features in precision mental health models. Incorporating more features, however, makes overfitting the data more likely. Thus, successful precision mental health requires an analytic approach that can handle high-dimensional data without becoming too overfit to generalize to new patients.

Applying machine learning approaches

Applying machine learning to precision mental health research can address these limitations of traditional analytic techniques. Machine learning is an alternative analytic technique that uses statistical algorithms trained on high-dimensional arrays (hundreds or even thousands) of features8. Flexibly considering many features simultaneously means these models can tap the tangled web of constructs that comprise complex clinical phenomena. Critically, this allows researchers to consider many features in the same model – unlike previous precision mental health research that was limited to considering very few features simultaneously. This high dimensionality across and within sets of related features is necessary to explain a high portion of variance in person-level treatment success.

Although machine learning models can handle very large numbers of features, this capacity comes at a cost, referred to as the “bias-variance trade-off"9. Too many features (particularly correlated features) yield unstable models that vary strongly based on the data used to develop them. High variance compromises model generalizability because a high variance (e.g., very flexible) model may not predict very accurately in new data. However, too few features (as well as other constraints on model characteristics) yield biased models that also do not predict well because they miss important predictive patterns and relationships. Machine learning uses various techniques (e.g., regularization, hyperparameter tuning, simultaneous consideration of many model configurations) to optimize this bias-variance trade-off to accommodate high-dimensional sets of features while reducing overfitting.8,9 Thus, machine

learning methods may allow us to build precision mental health models that both capture clinical complexity and generalize accurately to new data.

Finally, machine learning provides rigorous resampling techniques to fit and evaluate models in separate data9. Consequently, models generalize well to new patients because they are evaluated on out-of-sample prediction. In a simplest case, data can be divided into held-in and held-out samples. More sophisticated resampling techniques such as cross-validation involve dividing the data many times to create multiple held-in and held-out samples. These approaches offer significant advantages for 1) accurately selecting a best model among multiple model configurations, and 2) estimating how well that model will perform when applied to new data (e.g., new patients in a clinical setting). Applying machine learning can accomplish the goal in precision mental health of accurate, robust treatment selection for new patients.

Opportunities to improve equity in mental health treatment

Critically, the combination of machine learning methods with the precision mental health paradigm may be able to mitigate health disparities in our current treatment pipeline. Treatments are rarely designed or evaluated in diverse samples, though the NIH has launched some new initiatives to improve this disparity. This push becomes somewhat irrelevant, however, when we rely exclusively on treatments developed decades ago that were effective in homogeneous samples, and we discount treatments that were less effective within the research sample but may have worked well for people with characteristics or identities not well-represented in that sample. Unfortunately, the people who we failed to include in our treatment design and validation are also

those who are often disproportionately affected by these exact health conditions or face additional barriers to receive effective treatment.

These problems have, in some ways, been amplified by early work in precision medicine and precision mental health. Studies often begin with retrospective samples before bringing the models built in those samples to new patients. Thus, these models may be likely to fail for the same people for whom our initial treatments fail - the model can only be as good as the data with which it was developed. However, when this research is done thoughtfully, there are clear opportunities to address health disparities.

First, we can make a concerted effort to build treatment selection models using data from previously completed trials where there was good representation across as many marginalized characteristics as possible. This may include demographic variables such as sex, gender identity, race, or ethnicity, among others. Other important characteristics to consider might include income, access to insurance, and geographic region.

In some contexts, it would be problematic to use predictors that tap into constructs delineating marginalized identities such as race or socioeconomic status. For example, making decisions about who gets insurance (or doesn't) and who gets released earlier from prison (or doesn't) based on race would be not only inappropriate but also discriminatory. However, in the precision mental health landscape, we are not deciding *who* gets treatment. Rather, we are deciding *which* treatment to give a specific patient. Thus, we can take advantage of experiential or symptomatological differences as a result of characteristics such as race, ethnicity, sex, income, or comorbid health conditions to improve treatment outcomes.

An additional way in which this approach can help mitigate health disparities is through access. Many of the individual difference features that could help build treatment selection models may be easily measured via self-report, and dimensionality reduction approaches can limit the number of features that need to be assessed. Consequently, treatment selection models might require sparse assessment of only a handful of readily available items and could even be implemented online. In cases where treatments are available over-the-counter, completing the assessment remotely means that individuals without insurance or access to in-person medical care can still give themselves the best chance of treatment success.

Cigarette smoking as a critical precision mental health target

Public health importance of cigarette smoking

Cigarette smoking could benefit greatly from combining precision mental health and machine learning. Smoking remains an enormous public health burden. Tobacco is the number one cause of preventable death in the U.S. and accounts for more than 480,000 deaths annually10-12. Although rates of smoking have declined considerably, approximately 14% of U.S. adults continue to smoke daily or near-daily10-12. Cigarette smoking rates also remain much higher in potentially vulnerable populations including: people with chronic or severe mental illness (Baker ARCP); Native American and non-Hispanic Black individuals (CDC 2007); individuals who are economically and educationally disadvantaged (CDC 2007); people with other substance use disorders (Kelly 2012); people in the criminal justice system (Cropsey Eldridge Ladner 2004; Harrison et al 2019); people experiencing homelessness (); people who are insured through

Medicaid or uninsured (Jamal 2015); and individuals who identify as lesbian, gay, or bisexual (Jamal 2015).

Smoking cessation treatment

Despite the severity of the problem, treatment has had relatively limited reach (CITE Baker ARCP review). A survey of almost 16000 US adults who use cigarettes showed that the most commonly used strategies by far to quit were giving up cigarettes all at once and gradually cutting back, with a much smaller proportion using evidence-based treatments (CITE Carabello 2017).

For those who do get treatment, the best available smoking cessation treatments are modestly effective. The medications varenicline and combination nicotine replacement therapy (C-NRT) are consistently identified as the most effective options when combined with psychosocial counseling (Cahill et al., 2013), and guidelines recommend that clinicians consider either of these two medications first given their established efficacy (Fiore et al., 2008). These medications appear to be equally (though modestly) effective: a meta-analysis demonstrated comparable effectiveness rates (Cahill et al., 2013), and the first randomized controlled trial (RCT) directly comparing varenicline and C-NRT did not find a difference between them (Baker et al., 2016).

Typically, 6-month abstinence rates hover around 30-35% for these best smoking cessation medications combined with psychosocial counseling13,14. Treatment with an FDA-approved medication doubles the likelihood that an individual will quit successfully (Baker ARCP). These rates represent a best-case scenario in that clinical trial data involve treatment

regimens that are rigorously followed and optimized for adherence. Additionally, because several first-line (i.e., FDA-approved) smoking cessation treatments have comparable population-level effectiveness rates, population effectiveness alone cannot guide selection among smoking cessation treatments. These facts suggest a critical need for machine learning-assisted treatment selection in the cigarette smoking domain.

Opportunities for differential treatment selection

National clinical guidelines note that "there are no well-accepted algorithms to guide optimal selection" between any of the first-line medications (Fiore et al., 2008, p. 44). However, there may be reason to expect that some treatments may work better than others for a specific individual.

First, there is enormous heterogeneity among people who smoke cigarettes. Individuals may differ with respect to the etiology of their tobacco use disorder, the severity of their dependence and/or withdrawal symptoms, historical factors related to their tobacco use (e.g., age of first use, years smoking, number of previous quit attempts), and barriers to initiation and/or retention in smoking cessation treatments (Oliver & McClernon, 2017; J. Wang, Simons-Morton, Farhat, Farhart, & Luk, 2009; Zheng, Wiebe, Cleveland, Molenaar, & Harris, 2013). These factors that affect the development and course of their disorder could include demographic traits, personal medical history, and many other key individual difference characteristics. However, this heterogeneity is typically neglected when selecting among available treatments.

Second, smoking cessation medications have distinct pharmacological mechanisms of action at nicotinic acetylcholine receptors (nAChRs), which may affect how helpful they are for

different people. Nicotine replacement therapy (NRT) provides nicotine, a full agonist at nAChRs. Different NRTs provide nicotine differently. C-NRT consists of a nicotine patch and ad libitum nicotine lozenge use. The patch offers transdermal administration of a low, steady dose of nicotine. Some people who smoke cigarettes may rely on this low, steady nicotine level to replace nicotine from cigarettes. Lozenges provide oral administration of nicotine with more rapid onset, which could help individuals who need a quick boost during craving.

Other individuals who smoke may benefit from a medication like varenicline. In contrast to NRTs, varenicline is a partial agonist at nAChRs. Partial agonists have a pharmacological action that is somewhere between full agonists and antagonists, depending on the level of surrounding neurotransmitter. In the absence of a full agonist or endogenous neurotransmitter, partial agonists can act as a functional agonist with lower activity than a full agonist. In the presence of a full agonist (e.g., a cigarette) or endogenous neurotransmitter, however, they act as functional antagonists because their binding to the receptor limits the amount of binding from the full agonist and consequently reduces that response (Jordan & Xi, 2018; Lieberman, 2004).

Thus, varenicline may be more pharmacologically flexible than NRT medications: when an individual is not smoking, it can produce milder, nicotine-like effects; if an individual begins smoking again, it could block or reduce full agonist (cigarette nicotine) activity at the receptor. This would be expected to reduce the pharmacological effect of nicotine, likely reducing the behavioral pleasure of smoking (Cahill, Lindson-Hawley, Thomas, Fanshawe, & Lancaster, 2016). Although C-NRT has some behavioral flexibility built in (i.e., combination of slow, steady

dosing with faster-acting lozenges that can be used in response to internal states or environmental cues), it acts exclusively as a full agonist at nAChRs and cannot exert antagonist-like actions.

Third, features across several behavioral or environmental domains may also guide treatment selection, alone or in combination with medication mechanisms of action. For example, some cigarette smokers may have strong cravings with good self-monitoring. These characteristics may make treatments such as nicotine lozenges or gum more effective for those people because they are aware enough of their own craving to get a quick "hit" of nicotine when needed. Some smokers may be prone to side effects from a specific treatment, reducing adherence and subsequent likelihood of treatment success, though they may not have had the same adverse reactions to a different treatment. Environmentally, an individual who lives with other smokers may benefit from a partial agonist treatment like varenicline because any secondhand smoke would produce less effect. Any of these characteristics, among others, could powerfully inform treatment selection for cigarette smoking cessation. These examples illustrate the potential clinical benefit of using a precision mental health paradigm to inform treatment selection for smoking cessation. They also point to the value of analytic techniques that can incorporate complex interactions among features. Although these examples were selected because they are more intuitive, there are likely other unexpected ways that treatment success differs across people. Machine learning models are not limited to intuitive or theoretically derived features. Thus, machine learning may reveal unanticipated features that could meaningfully guide treatment selection for cigarette smoking cessation.

Finally, there is some evidence that individuals respond differently to different treatments. One large study that examined multiple medication-assisted quit attempts found that individuals who switched medications were more likely to quit than individuals who used the same medication again or who did not use a medication on the first quit attempt but added one at the second (Heckman et al., 2017). Relatedly, there is some evidence that re-treatment with the same medication as a previous, unsuccessful quit attempt is not effective; Gonzales and colleagues note that "abstinence rates are more than threefold lower for NRTs and twofold lower for bupropion" during re-treatment compared to initial treatment using the same medication (Gonzales et al., 2014, p. 391; Fiore et al., 2008; Tønnesen, Nørregaard, Säwe, & Simonsen, 1993). Additionally, clinical research related to other psychological and psychiatric disorders has demonstrated differential treatment benefit on an individual basis (e.g., antipsychotic medications for schizophrenia; Roussidis et al., 2013), suggesting it is worth investigating whether the same is true in smoking cessation.

Previous precision mental health & machine learning research

Purpose

Methods

Transparency & openness

We adhere to research transparency principles that are crucial for robust and replicable science. We reported how we determined the sample size, all data exclusions, all manipulations, and all study measures. We provide a transparency report in the supplement. Finally, our data, analysis

scripts, annotated results, questionnaires, and other study materials are publicly available (https://osf.io/qad4n/).

We preregistered our analyses to evaluate clinical benefit that relied on significance testing. The preregistration can be found on our OSF page (https://osf.io/qad4n/). For our Bayesian hierarchical generalized linear models, we followed guidelines from the tidymodels team (CITE) that we have followed in other published research from our laboratory (Wyant et al., in press). For all other analyses, we restricted many researcher degrees of freedom via cross-validation. Cross-validation inherently includes replication: models are fit on held-in training sets, decisions are made in held-out validation sets, and final performance is evaluated on held-out test sets.

Data

The data for this project came from a completed randomized controlled trial conducted by the University of Wisconsin (UW) Center for Tobacco Research and Intervention (CTRI)21. This trial compared the effectiveness of three cigarette smoking cessation treatments (varenicline, combination nicotine replacement therapy [C-NRT], and nicotine patch). Briefly, 1086 daily cigarette smokers were enrolled in Madison, WI, USA and Milwaukee, WI, USA. Exclusion criteria included contraindicated medical (e.g., severe hypertension) or psychiatric conditions (e.g., severe and persistent mental illness), current use of contraindicated medications, and pregnancy or unwillingness to use appropriate methods of contraception while taking a study

medication. Participants set a quit date with study staff and were enrolled for several weeks prior to the target quit date through at least 6 months following quitting smoking.

Treatment conditions

Participants were randomly assigned to receive 12 weeks of medication treatment plus 6 sessions of motivational and skill-training counseling per clinical guidelines (CITE FIORE GUIDELIINES). For varenicline, participants began medication use prior to their quit attempt, starting with 0.5 mg once daily for 3 days, followed by 0.5 mg twice daily for 4 days, and 1 mg twice daily for 3 days. They continued use of 1 mg twice daily for 11 weeks following their quit date except in response to adverse effects. For C-NRT or nicotine patch, participants began using the patch on their quit date, starting with 21 mg for 8 weeks, followed by 14 mg for 2 weeks, and 7 mg for 2 weeks. All individuals who received C-NRT were also instructed to use 5 lozenges per day (2 or 4 mg nicotine lozenges determined by time to first daily cigarette) for the full 12 weeks except in the case of adverse effects.

Individual difference characteristics

Participants were comprehensively assessed for individual differences characteristics prior to treatment randomization. These characteristics fall into several domains expected to relate to cigarette smoking cessation: tobacco-related (e.g., cigarettes per day), psychological (e.g., psychiatric diagnoses, distress tolerance), physical health (e.g., vital signs), social/environmental (e.g., living with another person who smokes), and demographic (e.g., age, sex). A detailed list of all available individual differences variables appears in Table X.

Abstinence outcome

Throughout study participation, participants were assessed for biologically confirmed, 7-day point-prevalence abstinence. Participants self-reported whether they had smoked over the past 7 days, and their report was biologically confirmed via exhaled carbon monoxide (CO). Participants were labeled as "abstinent" if their CO level was less than 6 parts per million (ppm; Baker et al., 2016). If participants self-reported smoking in the past 7 days, their CO level contradicted their self-report (i.e., CO level > 6 ppm), or biological confirmation could not be confirmed, participants were labeled as "smoking." Participants were assessed for abstinence periodically beginning 4 week post-quit through the end of their study participation. Our *primary prediction outcome* for our models was point-prevalence abstinence at 4 weeks post-quit.

AIM 1 analytic strategy: Model building

Feature engineering and dimensionality reduction

Feature engineering is the process of converting raw predictors into meaningful numeric and/or categorical representations (features) that improve model effectiveness (CITE kuhn feature engineering). A sample feature engineering script (i.e., tidymodels recipe) containing all feature engineering steps is available on our OSF study page (https://osf.io/qad4n/). Our generic feature engineering steps included: 1) imputing missing data (median imputation for numeric features, mode imputation for nominal and ordinal features); and 2) removing zero-variance features. Medians/modes for missing data imputation and identification of zero variance features were

derived from held-in (training) data and applied to held-out (validation and test) data (see Cross-validation section below).

We used a Yeo-Johnson transformation on all numeric variables to address distributional shape. Unordered categorical variables were dummy-coded. Ordered categorical variables (e.g., Likert scale items on self-report measures) could be ordinal scored (i.e., treated as numeric data) or dummy-coded. We also allowed treatment (dummy coded) to interact with all other features. Finally, all features were normalized as a requirement of the GLMNet algorithm.

Although machine learning methods can handle high-dimensional data, there is a cost to including a high ratio of features to observations ("*p* to *n* ratio"). Models were $p > n$ are possible with machine learning; however, models can become easily overfit and therefore not generalizable to new data. Thus, we used several dimensionality reduction approaches to reduce the number of features in my models. We used data-driven methods for dimensionality reduction including: removing highly correlated or near-zero variance features, and considering feature engineering approaches that reduced the number of overall features (e.g., ordinal scoring vs. dummy coding). The algorithm GLMNet also inherently conducts dimensionality reduction by penalizing model complexity such that features' predictive value must outweigh the cost of including an additional parameter in the model.

We also used several non-data-driven approaches for dimensionality reduction. First, we used domain knowledge to reduce dimensionality by removing features that lacked face validity for predicting abstinence or overlapped conceptually with other features. Second, we removed

variables that stood in contrast to the ultimate implementation goals (e.g., variables whose assessment required blood work or lab tests).

Model training and evaluation

Model configurations. All model configurations used the statistical algorithm Elastic Net Logistic Regression (GLMNet). This algorithm aligns with our primary goal of building a treatment selection model in several ways. First, it allows for explicit inclusion of interaction terms (i.e., including interactions between treatment and all other variables). This permits capturing multiple interactions that each account for a small portion of variance. Second, GLMNet performs a degree of dimensionality reduction because it penalizes model complexity; thus, a model using this algorithm may require assessing fewer features than are initially considered in the model. This characteristic aligns with our intention to implement this model in clinical practice, where highly burdensome assessments are impractical and thereby not feasible. Finally, linear models such as GLMNet are often more interpretable and transparent because they produce parameter estimates for the features in the model. Initial testing showed that GLMNet outperformed or performed comparably to models fit using several other well-established statistical algorithms (XGBoost, Random Forest). Thus, we had no reason not to prefer an algorithm that aligned well with our ultimate clinical goals.

Candidate model configurations differed across sensible values of the hyperparameters alpha and lambda (GLMNet tuning parameters). We also considered several outcome resampling techniques: no resampling, up-sampling, down-sampling, and the synthetic minority oversampling technique (SMOTE). These resampling techniques were used to create majority/

abstinence to minority/smoking ratios in the held-in training data ranging from 2:1 to 1:1 (natural rate in data ~3:1).

We considered several feature sets across model configurations. Feature sets could include either items (i.e., individual items within a self-report measure) or scales (i.e., total scale and sub-scale scores derived from items in a self-report measure). All other features (e.g., demographic variables) were included across model configurations.

Feature engineering steps also differed across model configurations regarding how ordinal data were scored. Specifically, ordinal data could be considered as numeric (e.g., 1 - 7) or dummy coded (e.g., 7 dummy code features for a 7-level variable). Numeric and unordered nominal data were treated identically across configurations.

Performance metric. Our primary performance metric for model selection and evaluation was area under the Receiver Operating Characteristic Curve (auROC) [CITE kuhnAppliedPredictiveModeling2018]. auROC indexes the probability that the model will predict a higher score for a randomly selected positive case (lapse) relative to a randomly selected negative case (no lapse). This metric was selected because it 1) combines sensitivity and specificity, which are both important characteristics for clinical implementation; and 2) is unaffected by class imbalance, which is important for comparing models with differing levels of class imbalance.

Cross-validation. We used nested cross-validation for model training, selection, and evaluation with auROC. Cross-validation allows for rigorous consideration of many model

configurations (i.e., combinations of feature sets, statistical algorithms, resampling techniques, and hyperparameters) and prioritizes performance in new data not used for model training.

Nested cross-validation uses two nested loops for dividing and holding out folds: an outer loop, where held-out folds serve as *test sets* for model evaluation; and inner loops, where held-out folds serve as *validation sets* for model selection. Importantly, these sets are independent, maintaining separation between data used to train the models, select the best models, and evaluate those best models. Therefore, nested cross-validation removes optimization bias from the evaluation of model performance in the test sets and can yield lower variance performance estimates than single test set approaches [CITE jonathanUseCrossvalidationAssess2000].

We used 1 repeat of 10-fold cross-validation for the inner loops and 3 repeats of 10-fold cross-validation for the outer loop. Best model configurations were selected using median auROC across the 10 *validation sets*. Final performance evaluation of those best model configurations used median auROC across the 30 *test sets*. We report median auROC for our best model configurations for each model (4-week and 26-week) in the test sets. For completeness, we also report auROCs for these models from the validation sets in the Supplement. In addition, we report other key performance metrics for the best full model configurations including sensitivity, specificity, balanced accuracy, positive predictive value (PPV), and negative predictive value (NPV) from the test sets [CITE kuhnAppliedPredictiveModeling2018].

Following model evaluation, we completed another round of 1 repeat of 10-fold cross-validation using the full dataset. A single best model configuration was selected using median auROC across the 10 held-out folds; importantly, model performance is used *only* for selection

and not evaluation in this phase. We fit our final model using this best model configuration in the full dataset, which was then used for clinical benefit analyses (below).

Evaluation of model performance

We used a Bayesian hierarchical generalized linear model to estimate the posterior probability distributions and 95% Bayesian confidence intervals (CIs) for auROC for the best models. Posterior probability is the likelihood of obtaining our results given our data. A posterior probability distribution around a given parameter (e.g., median auROC) allows us to assess the certainty of our results.

To estimate the probability that the 4-week model outperformed the 26-week model, we regressed the auROCs (logit transformed) from the 30 test sets for each model as a function of outcome (4-week vs. 26-week). Following recommendations from the tidymodels team [CITE kuhnTidyposteriorBayesianAnalysis2022], we set two random intercepts: one for the repeat, and another for the fold within repeat (folds are nested within repeats for 3x10-fold cross-validation). We report the 95% (equal-tailed) Bayesian CIs from the posterior probability distributions for our models' auROCs. If 95% Bayesian CIs do not include 0.5 (chance performance), we can conclude that the model performs better than chance. We also report 95% (equal-tailed) Bayesian CIs for the differences in performance associated with the Bayesian comparisons. If the 95% Bayesian CI around a difference in performance does not include 0, we can conclude that one model performs better than the other.

Bayesian analyses were accomplished using the tidyposterior [CITE kuhnTidyposteriorBayesianAnalysis2022] and rstanarm [CITE

goodrichRstanarmBayesianApplied2023] packages in R. Following recommendations from the rstanarm team and others [CITE rstudioteamRStudioIntegratedDevelopment2020; CITE gabryPriorDistributionsRstanarm2023], we used the rstanarm default autoscaled, weakly informative, data-dependent priors that take into account the order of magnitude of the variables to provide some regularization to stabilize computation and avoid over-fitting. Specifically, the priors were set as follows: residual standard deviation ~ normal(location=0, scale=exp(2)), intercept (after centering predictors) ~ normal(location=2.3, scale=1.3), the two coefficients for window width contrasts ~ normal (location=0, scale=2.69), and covariance ~ decov(regularization=1, concentration=1, shape=1, scale=1).

Feature importance with SHAP

We computed Shapley Values [CITE lundbergUnifiedApproachInterpreting2017] to provide a consistent, objective explanation of the importance of categories of features (based on EMA questions) across our three full models. Shapley values possess several useful properties including: Additivity (Shapley values for each feature can be computed independently and summed); Efficiency (the sum of Shapley values across features must add up to the difference between predicted and observed outcomes for each observation); Symmetry (Shapley values for two features should be equal if the two features contribute equally to all possible coalitions); and Dummy (a feature that does not change the predicted value in any coalition will have a Shapley value of 0).

We calculated Shapley values for the best model configuration from the 10 held-out folds in the final 1X 10-fold cross-validation used to select the final model. We used the DALEX

(CITE) and DALEXtra (CITE) packages in R, which provide Shapley values in log-odds units for binary classification models. These Shapley values estimate local importance (i.e., for each observation). To calculate global importance (i.e., across all observations), we averaged the absolute value of the Shapley values of each feature across observations. These local and global importance scores based on Shapley values allow us to answer questions of relative feature importance; however, these are descriptive analyses because standard errors or other indices of uncertainty for importance scores are not yet available for Shapley values.

AIM 2 analytic strategy: Evaluation of clinical benefit

Identify model-predicted best treatment

The preregistration for our analyses evaluating clinical benefit can be found on our OSF page (https://osf.io/qad4n/). Using our final model (best selected model configuration fit on all data), we calculated three predictions for each participant by substituting each treatment into the model inputs. Thus, there is one prediction per person per treatment. For example, an individual may have received varenicline in the original trial. We calculated their probability of abstinence using their data, and then we calculated two additional probabilities by substituting C-NRT and nicotine patch for varenicline. These substitutions affected the probabilities through any main effect of treatment and any interactions of treatment with other features.

The treatment that yields the highest model-predicted probability of abstinence is identified as that participant's "best" treatment. For example, among the three calculated probabilities for an individual, their probability of abstinence may be highest when C-NRT is

substituted in as their treatment. This would mean C-NRT is identified as the best treatment for that person.

Categorize treatment "matching"

Some participants' best treatment matched what they were randomly assigned in the original trial. Other participants may have received a sub-optimal treatment (i.e., what the model identified as their second-best or worst treatment based on calculated probabilities). Thus, participants' RCT-assigned treatment can be categorized by whether it "matched" their model-selected treatment.

Evaluate clinical benefit

Our primary analysis to evaluate the clinical benefit of our model-selected treatment compared the observed outcomes (i.e., abstinence vs. smoking from the original trial) for people who did or did not receive their best treatment. Treatment matching was thus a between-subjects predictor and was coded as 0.5 (TRUE) vs. −0.5 (FALSE).

We examined this effect over time at 4, 12, and 26 weeks by allowing the effect of treatment match to interact with time (i.e., week). Time was a within-subjects variable with three repeated measures for each participant. We treated time numerically, and we used a log transformation (base 2) to meet linearity assumptions. Thus, our model included treatment match, time, the interaction between treatment match and time, a by-subject random slope for time, and a by-subject random intercept.

We followed a mixed-effects modeling approach using the blme package (Chung et al., 2013). Specifically, we fit a partially Bayesian generalized linear model that uses regularizing priors to force the estimated random effects variance-covariance matrices away from singularity

(Chung et al., 2013). If the interaction between treatment match and time was significant, we planned to conduct follow-up tests of the simple effect of treatment match at each time point (week 4, week 12, and week 26) using general linear models.

We identified the main effect of treatment match *a priori* as our focal effect; however, we report all estimates, test statistics, *p*-values, and confidence intervals from all models.

Results

Sample characteristics

analysis sample inclusion criteria and final sample size here (full sample) descriptive statistics on demographics and maybe some tobacco-related characteristics tables

Model performance

We selected the best model configurations using auROCs from the *validation sets*. We report the median and IQR auROCs from the validation sets for these best model configurations in the Supplement. We evaluated these best model configurations using *test set* performance to remove optimization bias present in performance metrics from validation sets.

The median auROC across the 30 test sets for the 4-week model was XX (IQR = XX - XX, range = XX - XX). The median auROC across the 30 test sets for the 26-week model was XX (IQR = XX - XX, range = XX - XX). Additional performance metrics (not used for selection or primary evaluation) are reported in the Supplement.

We used the 30 test set auROCs to estimate the posterior probability distribution for the auROC of these models. The median auROCs from these posterior distributions were XX (4-

week model) and XX (26-week model). These values represent our best estimates for the magnitude of the auROC parameter for each model. The 95% Bayesian CI for the auROCs were relatively narrow and did not contain 0.5 (chance performance) for either the 4-week model [XX - XX] or the 26-week model [XX - XX]. Figure X displays posterior probability distributions for the auROC for the models by outcome.

Bayesian model comparisons

We used the posterior probability distributions for the auROCs to compare formally the 4- and 26-week models. The median increase in auROC for the 4- vs. 26-week model was XX (95% CI = [XX - XX]), yielding a probability of XX% that the 4-week model had superior performance. Figure X presents histograms of the posterior probability distributions for this model contrast.

Feature importance

Parameter estimates for retained variables

The glmnet algorithm offers two advantages with respect to understanding variable importance. First, the algorithm performs regularization using the hyperparameter alpha. This hyperparameter penalizes model complexity by shrinking parameter estimates and/or removing unimportant variables from the model entirely. Thus, variables are retained in the model only to the degree to which their contribution to performance outweighs the cost of having an additional parameter in the model. Consequently, we can review the retained predictor variables as a metric of feature importance.

The best 4-week model configuration retained XX features (best model configuration alpha = XX). Of the XX retained features, XX were treatment interaction variables, suggesting the importance of these interactions for prediction. These retained features require assessing XX unique items (e.g., multiple dummy variables are from a single item, an item is retained in an additive and interactive feature). Table X presents the retained features from the 4-week model configuration and their parameter estimates.

The best 26-week model configuration retained XX features (best model configuration alpha = XX). Of the XX retained features, XX were treatment interaction variables, suggesting the importance of these interactions for prediction. These retained features require assessing XX unique items (e.g., multiple dummy variables are from a single item, an item is retained in an additive and interactive feature). Table X presents the retained features from the 26-week model configuration and their parameter estimates.

Shapley values

Global importance (mean |Shapley value|) for features for each model appear in Panel A of Figure X. XX was the most important feature category across prediction outcomes. XX, XX, and XX were also globally important across models. XX, XX, and XX were the most relatively important treatment interaction variables.

Sina plots of local Shapley values (i.e., the influence of features on individual observations) for each model show that some features (e.g., XX, XX, XX) impact abstinence probability for specific individuals even if they are not globally important across all observations (Figure X, Panels B-C).

Clinical benefit

4 & 26 weeks

main effect of tx_match

tx_match X time interaction - follow-up simple effects if needed

supplemental aim 2 analyses: 1/2/3 tx rank?

Discussion

References

Bibliography