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 [derubeisHistoryCurrentStatus2019; inselNIMHResearchDomain2014; bickmanAchievingPrecisionMental2016]. 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 [bickmanImprovingMentalHealth2020]. 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 [bickmanAchievingPrecisionMental2016; bickmanImprovingMentalHealth2020].

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 slow. These costs may also produce a treatment that is no better than

existing treatments (e.g., [weiszArePsychotherapiesYoung2019]), 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 been tied closely to genetic factors and thus has grown rapidly with the ascendancy of advanced genetic methods such as genome-wide association studies, polygenic scores, and functional annotation [wrayResearchReviewPolygenic2014; bogdanPolygenicRiskScores2018; kranzlerPrecisionMedicinePharmacogenetics2017; huLeveragingFunctionalAnnotations2017; finucanePartitioningHeritabilityFunctional2015]. Meaningful progress towards precision treatments has been made in the cancer domain; for example, some chemotherapy drugs that are not effective at a population level have been shown to help individuals with specific non-small-cell lung carcinoma tumor mutations [rosellErlotinibStandardChemotherapy2012]. Perhaps unsurprisingly, these advances have been propelled by significant funding - cancer research has received far and away the most NIH funding over the past 25 years [kranzlerPrecisionMedicinePharmacogenetics2017].

Within precision mental health, an early research 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 severity [projectmatchresearchgroupProjectMATCHMatching1993;

projectmatchresearchgroupMatchingAlcoholismTreatments1998]. Many researchers have followed in their footsteps as the understanding has grown that neither mental health diagnoses nor treatments are one-size-fits-all [derubeisHistoryCurrentStatus2019]. This research has primarily focused on selecting among treatments for depression (e.g., [derubeisPersonalizedAdvantageIndex2014; webbPersonalizedPredictionAntidepressant2019]; for review, see [cohenTreatmentSelectionDepression2018]).

Despite these opportunities for advances, however, precision mental health research has progressed with limited success [kesslerPragmaticPrecisionPsychiatry2021]. 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 differently [derubeisHistoryCurrentStatus2019]. However, 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 [kesslerPragmaticPrecisionPsychiatry2021; inselNIMHResearchDomain2014]. This idea is 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 [bogdanPolygenicRiskScores2018; chenPathwaysPrecisionMedicine2018;

wrayResearchReviewPolygenic2014]. 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; [bickmanImprovingMentalHealth2020]). 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 [jonathanUseCrossvalidationAssess2000]. 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 may be able to address these limitations of traditional analytic techniques to advance precision mental health goals [bickmanAchievingPrecisionMental2016; dwyerMachineLearningApproaches2018; maceachernMachineLearningPrecision2021; mooneyBigDataPublic2018]. Machine learning is an alternative analytic technique that uses statistical algorithms trained on high-dimensional arrays (hundreds or even thousands) of features [jamesIntroductionStatisticalLearning2013; kuhnAppliedPredictiveModeling2018; ngMachineLearningYearning2018]. 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 [maceachernMachineLearningPrecision2021]. 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" [jamesIntroductionStatisticalLearning2013; ngMachineLearningYearning2018]. 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 [jamesIntroductionStatisticalLearning2013; kuhnAppliedPredictiveModeling2018; ngMachineLearningYearning2018; mooneyBigDataPublic2018]. 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 data [jamesIntroductionStatisticalLearning2013]. 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 [krstajicCrossvalidationPitfallsWhen2014; jonathanUseCrossvalidationAssess2000]. 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 [maceachernMachineLearningPrecision2021]. Treatments are rarely designed or evaluated in diverse samples; input directly from patient shareholders is incorporated even less often. Although the NIH has launched some new initiatives to improve effort in these areas, when it comes to treatments, we still rely almost exclusively on treatments developed decades ago that were effective in homogeneous samples. 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 [moralesCallActionAddress2020; barksdaleInnovativeDirectionsAdvance2022; officeofthesurgeongeneralusMentalHealthCulture2001; jacobsonDigitalTherapeuticsMental2022].

In some ways, these problems have 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 [aldridgeResearchTrainingRecommendations2019]. 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 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 discriminatory (e.g., [farayolaEthicsTrustworthinessAI2023]). 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 across vulnerable subpopulations.

An additional way in which this approach can help mitigate health disparities is through access. Access to treatment is a known barrier in mental healthcare and a contributing factor driving mental healthcare disparities [jacobsonDigitalTherapeuticsMental2022]. Many individual difference features that could help build treatment selection models may be easily measured via self-report, and dimensionality reduction approaches employed within machine learning 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 annually [nationalcenterforchronicdiseasepreventionandhealthpromotionusofficeonsmokingandhealthHealthConsequences schlamInterventionsTobaccoSmoking2013; corneliusTobaccoProductUse2020]. Compared to people who have never smoked, individuals who smoke have two- to three-fold likelihood of death across causes and lose over a decade of life expectancy [jhaprabhat21stCenturyHazardsSmoking2013].

Although rates of smoking have declined considerably, approximately 14% of U.S. adults continue to smoke daily or near-daily [corneliusTobaccoProductUse2020]. Additionally, cigarette smoking rates remain much higher in potentially vulnerable populations. This includes people with chronic or severe mental illness [bakerSmokingTreatmentReport2021]; Native American and non-Hispanic Black individuals [jamalCurrentCigaretteSmoking2015a; corneliusTobaccoProductUse2020]; individuals who are economically and educationally disadvantaged [jamalCurrentCigaretteSmoking2015a; corneliusTobaccoProductUse2020]; people

with other substance use disorders [kellyPrevalenceSmokingOther2012]; people in the criminal justice system [cropseySmokingFemalePrisoners2004; harrisonCigaretteSmokingMental2020]; people experiencing homelessness [baggetttravisp.TobaccoUseHomeless2013; soarSmokingAmongstAdults2020]; people who are insured through Medicaid or uninsured [jamalCurrentCigaretteSmoking2015a; corneliusTobaccoProductUse2020]; and individuals who identify as lesbian, gay, or bisexual [jamalCurrentCigaretteSmoking2015a; corneliusTobaccoProductUse2020].

Smoking cessation treatment

Despite the severity of the problem, treatment has had relatively limited reach [bakerSmokingTreatmentReport2021]. 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 [caraballoQuitMethodsUsed2017]. Although some evidence suggests that 40-50% of former smokers quit on their first serious attempt, best estimates suggest it takes on average 30 or more attempts to quit successfully [chaitonEstimatingNumberQuit2016]. Individuals are less likely to achieve sustained abstinence when they have one or more failed quit attempts within the past year [partosQuittingRollercoasterHow2013], highlighting the need for individuals to receive effective treatment as early as possible.

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 [cahillPharmacologicalInterventionsSmoking2013; schlamInterventionsTobaccoSmoking2013; bakerSmokingTreatmentReport2021; rigottiTreatmentTobaccoSmoking2022]. Guidelines recommend that clinicians consider either of these two medications first given their established efficacy [fioreClinicalPracticeGuideline2008]. These medications appear to be equally (though modestly) effective: a meta-analysis demonstrated comparable effectiveness rates [cahillPharmacologicalInterventionsSmoking2013], and the first randomized controlled trial (RCT) directly comparing varenicline and C-NRT did not find a difference between them [bakerEffectsNicotinePatch2016].

Typically, 6-month abstinence rates hover around 30-35% for these best smoking cessation medications combined with psychosocial counseling [cahillPharmacologicalInterventionsSmoking2013; fioreClinicalPracticeGuideline2008]. Treatment with an FDA-approved medication doubles the likelihood that an individual will quit successfully [bakerSmokingTreatmentReport2021]. 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 [fioreClinicalPracticeGuideline2008, 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 [oliverPrecisionMedicineAddiction2017; wangSociodemographicVariabilityAdolescent2009; zhengIdiographicExaminationDaytoDay2013]. 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 [cahillNicotineReceptorPartial2016]. 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 [jordanDiscoveryDevelopmentVarenicline2018; liebermanDopaminePartialAgonists2004].

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 [cahillNicotineReceptorPartial2016]. 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 because

they are aware enough of their own craving to get a quick "hit" of nicotine when needed. In contrast, although C-NRT is largely more effective than single NRT (e.g., nicotine patch alone) [cahillPharmacologicalInterventionsSmoking2013], there may be individuals for whom a multi-component treatment is overwhelming, thus reducing adherence and ultimate effectiveness. 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 [heckmanEffectivenessSwitchingSmokingCessation2017]. Relatedly, there is some evidence that re-treatment with the same medication as a previous, unsuccessful quit attempt is not effective [fioreClinicalPracticeGuideline2008; tonnesenRecyclingNicotinePatches1993] Indeed, Gonzales and colleagues found 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 [gonzalesRetreatmentVareniclineSmoking2014, p. 391]. A pilot trial showed that treatment adherence improves when individuals are given the opportunity to sample various NRT medications pre-quit [cropseyPilotTrialVivo2017], and meta-analyses show only smokers who are highly dependent may benefit from 4 mg (vs. 2 mg) nicotine gum [lindsonDifferentDosesDurations2019]. These data suggest differential preferences and even differential effectiveness despite a shared pharmacological mechanism of action across NRT medications. 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 [roussidisReasonsClinicalOutcomes2013], psychosocial interventions for depression [cohenTreatmentSelectionDepression2018]), suggesting it is worth investigating whether the same is true in smoking cessation.

Previous precision mental health & machine learning research

In recognition of the need for improved treatment effectiveness and the potential for personalized treatment approaches, many smoking cessation researchers have pursued precision mental health research in recent years.

There is a reasonable body of evidence identifying "prognostic" factors [cohenTreatmentSelectionDepression2018] related to smoking cessation. This research aims to predict who will or will not be able to quit successfully [laiDevelopmentMachineLearning2021b; kaufmannRateNicotineMetabolism2015; kayeSearchingPersonalizedMedicine2020a; piperPrecisionSmokingCessation2017a; issabakhshMachineLearningApplication2023; etterPredictingSmokingCessation2023]. Similarly, there is research using machine learning that classifies individuals who smoke (vs. individuals who do not; [pariyadathMachineLearningClassification2014a]). A systematic review identified that predictors of smoking cessation outcomes span many categories: economic variables, environmental variables, sociodemographic variables, engagement in treatment, physical health variables, psychological variables, neurocognitive factors, biomarkers, and factors related to smoking history and severity [bickelPredictorsSmokingCessation2023].

Related research in this area seeks to understand who will succeed using a single treatment. Massago and colleagues (2024) examined medical records of individuals completing cognitive-behavioral treatment (CBT) and built a machine learning model to predict who is most likely to quit smoking using CBT. They note that this model "can be used to establish priorities when the demand is higher than the capacity" [massagoApplicabilityMachineLearning2024, p. 10]. Other research has used tree-based machine learning models to show that delay discounting tasks can differentiate individuals who respond to group CBT from individuals who do not [coughlinMachineLearningApproachPredicting2020a].

This research investigating prognostic factors is critical for understanding mechanisms that underlie smoking cessation success. Progress in this area may allow us to identify processes that improve the likelihood of quitting successfully, and we may be able to intervene in some of these processes prior to a quit attempt or improve design of preventative interventions. For example, individuals are less likely to succeed in a quit attempt when cigarettes are available and when confidence is low [bickelPredictorsSmokingCessation2023]. Unsurprisingly, key steps in pre-cessation and cessation counseling include getting rid of cigarettes and increasing confidence to quit. As more prognostic factors are identified, we can improve overall treatment effectiveness by targeting key predictors of success.

Although models that identify prognostic factors are important, they do not offer an actionable way forward to select among treatment options. Even research that informs us as to who might succeed within a specific treatment has limited utility for treatment selection: what do we do for a patient who is predicted not to succeed using that treatment? If different single-treatment prognostic models offer conflicting recommendations, how should a clinician or patient proceed? To *select among* treatments, we need "prescriptive" predictors [cohenTreatmentSelectionDepression2018], or predictors that determine which treatment to prescribe. This type of predictor is what allows us to select the best treatment for a given patient.

Some research has begun to investigate factors that can allow selection among treatment options, particularly genetic factors and biomarkers (for review, see [chenPathwaysPrecisionMedicine2018]). Chen and colleagues found that variants in the cholinergic receptor nicotinic alpha 5 subunit (CHRNA5) predict differential treatment success

among Black individuals who smoke: C-NRT was more effective for people with certain genotypes, whereas varenicline was more effective for people with other genotypes [chenGeneticVariantCHRNA52020]. Many researchers have investigated the role of the nicotine metabolite ratio (NMR), a phenotypic biomarker for nicotine metabolism that is easier to test than genotypes. Overall, it appears that slow metabolizers report greater cessation success with NRT medications, whereas varenicline is more effective for fast metabolizers [shahabDoesNicotineMetabolite2019; schnollNicotineMetabolicRate2009; glatardAssociationNicotineMetabolism2017; chenowethNicotineMetaboliteRatio2016; UseNicotineMetabolite], though a review by Siegel and colleagues notes that "real-world evaluations of NMR to personalize treatment for smoking cessation have produced mixed findings so far" ([siegelUseNicotineMetabolite2020], p. 265).

This research has exciting implications for selecting among treatments. However, using biologic and genetic factors comes with downsides when considering implementation. Though the necessary technology is improving, the testing required to collect these data is neither widely available nor accessible [maceachernMachineLearningPrecision2021]. Moreover, using these data to select treatments is likely to exacerbate rather than mitigate existing disparities in mental healthcare. Using genetic and biological factors is likely to favor privileged individuals who have insurance that covers specialized care and specialty testing. Some have noted that biomarker testing for NMR is more affordable than genetic testing; however, there remain methodological challenges associated with establishing cut-off points for slow vs. fast metabolizers, particularly among vulnerable subpopulations for whom there are no NMR data available yet who experience

greater smoking rates or associated risks (e.g., Native American/American Indian people, pregnant women, people who drink heavily or have other substance use disorders, people with serious mental illness, and individuals who identify as LGBTQ; [siegelUseNicotineMetabolite2020]).

Other research has investigated non-biological factors as potential moderators of treatment effectiveness. Kaye and colleagues (2020) examined whether varenicline might be more effective than C-NRT among individuals who smoke and binge drink; however, varenicline's effects did not vary as a function of drinking status [kayeSearchingPersonalizedMedicine2020a]. Piper and colleagues conducted a series of studies with the goal of advancing precision mental health for smoking cessation. They used a factorial design to identify ideal treatment combinations at the group level (e.g., pre-quit medication, pre-quit counseling, counseling modality, cessation medications; CITE ORIGINAL PIPER FACTORIAL STUDY). They followed up by exploring moderators of intervention main effects and interactions and found that psychiatric history moderated some treatment effects [piperPrecisionSmokingCessation2017a]. They also looked at whether various treatment components affected their purported treatment mechanisms to help elucidate how these treatments may be producing their effects [piperPrecisionSmokingCessation2017]. These studies took advantage of largely self-report data, making any resulting models more clinically implementable. However, each study examined only a single moderator or examined each moderator in a separate model. These analyses were ideally suited for clearly identifying

moderators and explaining mechanisms, but they were less well-suited for finding high-dimensional sets of features that can predict differential treatment responses.

Purpose

The goal of this project was to produce a model that can serve as a decision-making tool to select among medication treatments for smoking cessation. We used machine learning analytic techniques to build a model that can take advantage of many features simultaneously while maintaining generalizability to new patients. We incorporated easy-to-collect self-report data as our primary model inputs such that any resulting model will be poised for accessible, equitable clinical implementation.

Specifically, in this project we built a machine learning model to predict treatment success 4 weeks post-quit for people who smoke who received one of three cigarette smoking cessation treatments from a previously completed comparative effectiveness trial [bakerEffectsNicotinePatch2016]. This model was used to calculate probabilities of treatment success for each treatment to guide selection of the best treatment for any specific individual. We evaluated the clinical benefit of our model in these retrospective data while using resampling approaches that ensure our model offers benefit for treatment selection for new patients.

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/). These analyses also closely followed published research evaluating treatment selection models [derubeisPersonalizedAdvantageIndex2014]. For our Bayesian hierarchical generalized linear models, we followed guidelines from the tidymodels team [kuhnTidyposteriorBayesianAnalysis2022] that we have followed in other published research from our laboratory [wyantMachineLearningModels2023]. 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) [bakerEffectsNicotinePatch2016]. 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 [fioreClinicalPracticeGuideline2008]. 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; [bakerEffectsNicotinePatch2016]). 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."

Our *primary prediction outcome* for our models was point-prevalence abstinence at 4 weeks post-quit. Later assessments of point-prevalence abstinence (12 weeks [end-of-treatment], 6 months) were used for clinical benefit analyses (see below) and supplemental analyses.

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 [kuhnFeatureEngineeringSelection2019]. 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

Our classification models predicted 4-week point-prevalence abstinence. We also built models predicting 26-week (6 month) point-prevalence abstinence for secondary analyses; all model fitting and evaluation procedures were identical.

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 many interactions that each account for a small portion of variance; as discussed, this seems critical to capturing the complexity of clinical phenomena. 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. Several recent reviews have noted interpretability as a potential barrier to using machine learning approaches for clinical and public health goals; consequently, we aimed to prioritize interpretability [maceachernMachineLearningPrecision2021;

mooneyBigDataPublic2018; cohenTreatmentSelectionDepression2018]. 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) [kuhnAppliedPredictiveModeling2018; youngstromPrimerReceiverOperating2014]. 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. 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 [jonathanUseCrossvalidationAssess2000]. Specifically, we used nested cross-validation for model training, selection, and evaluation with auROC [krstajicCrossvalidationPitfallsWhen2014].

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 [jonathanUseCrossvalidationAssess2000; krstajicCrossvalidationPitfallsWhen2014].

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 in the test sets. For completeness, we also report auROCs from the validation sets in the Supplement. In addition, we report other key performance metrics for the best model configurations including sensitivity, specificity, balanced accuracy, positive predictive value (PPV), and negative predictive value (NPV) from the test sets [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 [Krstajic et al., 2014]. This best selected model configuration 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 of different performance across models, we regressed the auROCs (logit transformed) from the 30 test sets as a function of model outcome (4-week model vs. 26-week model). Following recommendations from the tidymodels team [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 [kuhnTidyposteriorBayesianAnalysis2022] and rstanarm [goodrichRstanarmBayesianApplied2023] packages in R. Following recommendations from the rstanarm team and others [rstudioteamRStudioIntegratedDevelopment2020; 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=XX, scale=exp(XX)), 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 [lundbergUnifiedApproachInterpreting2017] to provide a consistent, objective explanation of the importance of features. 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 by conducting leave-one-out cross-validation (LOOCV) using the final, best selected model configuration. LOOCV works identically to *k*-fold cross-validation described above, where $k = N$ (sample size) such that each test set consists of a single held-out participant. Thus, we fit $N = 1086$ models, where each participant served as the test set once for a model fit with the other 1085 participants. This allowed us to calculate Shapley values in held-out data, while ensuring our model stayed as close as possible to the final model (using the full dataset) that we would disseminate going forward.

We used the DALEX and DALEXtra packages [biecekDALEXExplainersComplex2018] 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

We followed closely methods described and used previously for evaluating the potential clinical utility of treatment selection models [derubeisPersonalizedAdvantageIndex2014]. We also preregistered our analyses to evaluate clinical benefit; the preregistration can be found on our OSF page (https://osf.io/qad4n/).

Identify model-predicted best treatment

As we did for feature importance, we conducted LOOCV such that we fit 1086 models using the best model configuration (selected in 1x-10-fold cross-validation) with each participant held-out from model fitting once. For each participant, we fit a model using the other 1085 participants and made predictions for the single held-out participant. This ensured that we were making predictions for a new patient (i.e., one that our model had not seen) to match most closely how this model would be ultimately implemented.

We calculated three predictions for each participant by substituting each treatment into the model inputs. Thus, there is one prediction per person per treatment. These substitutions affected the model's predicted 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, an individual may have received varenicline in the original trial. We calculated their probability of abstinence using their data (i.e., with varenicline as the "treatment" feature), and then we calculated two additional probabilities by substituting C-NRT and nicotine

patch for varenicline. Among these three predicted probabilities, 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 individual.

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.

For example, the individual described above received varenicline in the original trial, but their model-predicted probability of abstinence was highest when C-NRT was substituted in as their treatment. This participant's best treatment did not match their trial treatment, so they would be labeled as "unmatched."

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 (i.e., matched or unmatched). Treatment matching was thus a between-subjects predictor and was coded as 0.5 (TRUE, "matched") vs. −0.5 (FALSE, "unmatched").

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 to meet linearity assumptions. We preregistered using a log transformation with base e; however, due to issues of convergence, we switched to log base 2.

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 [chungNondegeneratePenalizedLikelihood2013a]. 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 [chungNondegeneratePenalizedLikelihood2013a]. 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 [krstajicCrossvalidationPitfallsWhen2014].

The median auROC across the 30 test sets for the 4-week model was 0.682 (IQR = 0.654 - 0.712, range = 0.589 - 0.797). The median auROC across the 30 test sets for the 26-week model was 0.645 (IQR = 0.605 - 0.672, range = 0.513 - 0.762). These values are comparable to model performance from extant literature predicting smoking cessation using machine learning (e.g., auROC = 0.660 [Lai et al 2021]). 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 0.687 (4-week model) and 0.639 (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 [0.666 - 0.707] or the 26-week model [0.617 - 0.661]. 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 0.047 (95% CI = 0.030 - 0.065), yielding a probability of 100% that the 4-week model had superior performance. Figure X presents histograms of the posterior probability distribution for this model contrast.

Feature importance

Parameter estimates for retained variables

The glmnet algorithm offers two advantages with respect to understanding variable importance. First, as a linear model, it outputs parameter estimates for each feature. Second, the algorithm performs regularization using the hyperparameter alpha. This hyperparameter penalizes model complexity by shrinking parameter estimates and/or removing unnecessary or highly correlated features from the model entirely. Thus, features 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 features as a metric of feature importance.

Table X presents the retained features from the best 4-week model configuration and their parameter estimates. This model retained 128 features (best model configuration alpha = 0.1, item feature set). Of the 128 retained features, 56 were treatment interaction features, suggesting the importance of these interactions for prediction. To perform treatment selection, only interactive features would need to be assessed, as features that increase or decrease probability

magnitude equally across all three treatments do not help with differential prediction. Consequently, implementing this model for treatment selection would require assessing only 37 unique items (e.g., multiple dummy variables are from a single item, a feature interacts with two levels of treatment).

Table X presents the retained features from the best 26-week model configuration and their parameter estimates. This model retained 38 features (best model configuration alpha = 0.3, scale feature set). Of the 38 retained features, 10 were treatment interaction variables, suggesting the importance of these interactions for prediction. To implement this model for treatment selection, this model would require assessing only 13 unique items (some retained items constituted scale scores and required assessing multiple items to calculate).

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 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. XX was globally important for only the 4-week model, whereas XX was globally important for only the 26-week model.

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-week model

There was a significant fixed effect of treatment matching on abstinence (OR = , $z = 5.640$, $p < 0.001$). Individuals who received their model-predicted best treatment were more likely to be abstinent. There was also a significant fixed effect of time (OR = , $z = -9.948$, $p < 0.001$) such that the probability of abstinence declined over time.

There was not a significant interaction between treatment matching and time ($p = 0.830$). However, calculating and interpreting interactions in logistic models is not straightforward because significance can differ based on the link function used [mandicInteractionTermsNonlinear2012; collinsOptimizationBehavioralBiobehavioral2018]. Consequently, we conducted simple effects analyses of the effect of treatment matching at each time point . This allowed us to characterize our results more fully and to understand our effects in their original probability terms.

There was a significant fixed effect of treatment matching on abstinence at 4 weeks (OR = , $z =$, $p =$) and at 12 weeks (OR = , $z =$, $p =$). The effect of treatment matching was no longer significant by the 26-week follow-up assessment ($p =$). Figure X shows the mean abstinence rate by treatment matching at each time point.

26-week model

Discussion

References

Bibliography