Machine learning-assisted treatment selection for smoking cessation

Gaylen E Fronk[aff-1] and John J. Curtin[aff-1]

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

Abstract

Precision mental health seeks to select the right treatment for a patient given personal characteristics. The purpose of this project was to build a machine learning model that could select among first-line medication treatments for cigarette smoking. We used data from a previously completed comparative effectiveness trial in which participants were richly characterized at baseline before being randomly assigned to varenicline, combination nicotine replacement therapy, or nicotine patch. We built a model predicting treatment success (abstinent vs. smoking) using baseline characteristics and their interactions with treatment. Models were fit, selected, and evaluated using nested cross-validation and the performance metric area under the receiving operator characteristic curve (auROC). Our best models had a median auROC of 0.69 in held-out test sets. We used this model to calculcate probabilities of smoking cessation success for each participant on each of the three treatments to identify their model-predicted best treatment. Individuals who received their model-predicted best treatment during the original trial were more likely to quit successfully than individuals who did not (OR = 1.851, p = 0.004). This project produces a clinically implementable treatment selection model to assist people quitting cigarette smoking.

Keywords: Substance use disordersPrecision mental healthCigarette smokingMachine learningTreatment selection

Introduction

There is a critical flaw in how we traditionally select treatments for patients. Selection primarily relies on population-level effectiveness, but 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 *on average*, it may be that treatment B is markedly more effective for a *specific* patient. Moreover, when our treatments are not as effective as they need to be, the pipeline for developing new treatments to fill those gaps is slow and costly: On average, new drug development takes 9-12 years and costs hundreds of millions of US dollars (Dickson & Gagnon, 2009).

Precision medicine may offer a solution. Precision medicine, sometimes called personalized medicine, seeks to guide treatment selection using individual difference characteristics that are likely to predict treatment success for each patient (Bickman, 2020). Successful precision medicine 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 (Bickman, 2020; Bickman et al., 2016). Precision medicine may also be more resource-efficient. Existing treatments can be optimized by directing them to the *right* patients - positioning this approach for immediate impact to patients. We can also use precision medicine to guide treatment development more strategically as we identify factors that promote treatment success or sub-populations for whom no existing treatments are optimal. This ensures that the time and cost that go into new treatment development are reserved for specific niches of need.

Researchers have pursued precision medicine for decades. Emphasis on personalizing treatments has been tied closely to genetic factors and has grown rapidly with the ascendancy of advanced genetic methods such as genome-wide association studies, polygenic scores, and functional annotation (Bogdan et al., 2018; Finucane et al., 2015; Hu et al., 2017; Kranzler et al., 2017; Wray et al., 2014). Perhaps the most meaningful progress towards precision treatments has been made in the cancer domain (Kranzler et al., 2017). 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 (Rosell et al., 2012).

Precision mental health

Precision mental health is the application of the precision medicine paradigm to mental health conditions (Bickman et al., 2016; DeRubeis, 2019; Insel, 2014). There is a critical need for improved mental health treatments. In the U.S. in 2022, over 59 million individuals aged 12 or older had a mental health disorder, and over 48 million people had a substance use disorder (Substance Abuse and Mental Health Services Administration, 2023). Mental health disorders are leading causes of disability and death (Centers for Disease Control and Prevention (CDC), n.d.; Whiteford et al., 2013). Psychiatric disorders also account for enormous economic burden; indeed, some estimates suggest higher than even serious medical disorders like cancer (Substance Abuse and Mental Health Services Administration (US) & Office of the Surgeon General (US), 2016; Trautmann et al., 2016). Additionally, mental healthcare is plagued by disparities related to race, ethnicity, geographic region, and socioeconomic status: Vulnerable sub-populations are more likely to have higher rates of mental health and substance use disorders and more difficulty accessing treatment (Barksdale et al., 2022; Jacobson et al., 2022; Morales et al., 2020; Office of the Surgeon General (US) et al., 2001; Substance Abuse and Mental Health Services Administration, 2023).

Treatments for mental health conditions are also usually no more than moderately effective, and many treatments for the same disorder can be quite comparable, making it difficult to select among them. For example, prolonged exposure therapy and cognitive processing therapy are two gold-standard treatments for post-traumatic stress disorder, and comparative trials have shown similar effectiveness (Lewis et al., 2020). Similarly, a recent review of anxiety treatments concluded that treatment effectiveness has not changed over decades despite the development of many new treatments in that period (Weisz et al., 2019). Among medication treatments, there is equal effectiveness across common antidepressants including sertraline, fluoxetine, and escitalopram (Adjei & Ali, 2022).

Like precision medicine, many researchers have pursued precision mental health. An early research example comes from the substance use disorder domain: The Project MATCH Research Group attempted to match people with alcohol use disorder to a particular treatment based on individual differences (Project Match Research Group, 1993; 1998). They assigned people to one of three treatments and evaluated whether any of ten patient attributes (e.g., gender, social support, symptom severity) interacted with treatment success. Despite a large body of evidence supporting the importance of characteristics such as these for predicting alcohol-related outcomes, they did not

find that any interactions had an impact on drinking. Although Project MATCH did not succeed in its goal of matching patients to treatments, it set the stage for future precision mental health research. Many researchers have followed in their footsteps as the understanding has grown that neither mental health diagnoses nor treatments are one-size-fits-all (DeRubeis, 2019). A large proportion of this research has focused on selecting among treatments for depression (e.g., (DeRubeis et al., 2014; Webb et al., 2019); for review, see (Cohen & DeRubeis, 2018)).

Despite the abundance of research, much less progress has been made in precision mental health than in precision medicine (Bickman, 2020; Bickman et al., 2016; Kessler & Luedtke, 2021; Kranzler et al., 2017; Oliver & McClernon, 2017). There may be several reasons for this. First, mental health disorders generally lack the "rigorously tested, reproducible, clinically actionable biomarkers" that have paved the way for progress in precision medicine (Insel, 2014; Kranzler et al., 2017). This means that we must either identify and validate these kinds of biomarkers for mental health disorders or look to other domains such as demographic, psychological, and environmental factors (Bickman, 2020).

Second, many factors influence heterogeneous, complex clinical phenomena like mental health diagnoses and treatment success (Feczko & Fair, 2020). Thus, any single feature (i.e., predictor variable) cannot account for more than a small portion of the variance in treatment success (Insel, 2014; Kessler & Luedtke, 2021). 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 (Bogdan et al., 2018; Chen et al., 2018; Wray et al., 2014). Consider that, although Project MATCH did not find any interactions among treatments and patient characteristics, these were each tested in a separate model. It may be that the factors they tested *were* important - but none explained sufficient variance in a clinical outcome *on its own*.

Precision mental health efforts so far, however, have largely focused on personalizing treatments by identifying a single factor that divides individuals within a diagnostic category into subgroups that can be treated differently (DeRubeis, 2019). Given the known heterogeneity in mental health diagnoses, groups differentiated on a single factor are not homogeneous - group-level tailoring may not be sufficiently granular for mental health disorders. It is perhaps unsurprising that models that consider only one or a small handful of features - which also limits considering concurrently features across categories - have failed to capture the real-world complexity underlying clinical phenomena like treatment success.

Additionally, because precision mental health models are typically developed and evaluated in the same sample, the models may become very overfit to that sample (Jonathan et al., 2000). This problem is exacerbated in precision mental health because sample sizes in psychological research have remained relatively small despite recommendations to increase sample size (Marszalek et al., 2011). Consequently, precision mental health models do not generalize well to new patients. Clinical implementation of these models requires that they provide accurate, patient-level recommendations about treatment selection for *new* patients.

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 to advance precision mental health goals (Bickman et al., 2016; Dwyer et al., 2018; MacEachern & Forkert, 2021; Mooney & Pejaver, 2018). Machine learning is an alternative analytic technique that uses statistical algo-

rithms trained on high-dimensional arrays (hundreds or even thousands) of features (James et al., 2013; Kuhn & Johnson, 2018; Ng, 2018). Flexibly considering many features simultaneously means these models can tap the tangled web of constructs that comprise complex clinical phenomena like treatment success. 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 (MacEachern & Forkert, 2021). 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" (James et al., 2013; Ng, 2018). 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) within cross-validation to optimize this bias-variance trade-off to accommodate high-dimensional sets of features while reducing overfitting (James et al., 2013; Kuhn & Johnson, 2018; Mooney & Pejaver, 2018; Ng, 2018). In these ways, machine learning methods may allow us to build precision mental health models that both capture clinical complexity and generalize accurately to new data.

In addition to affecting the bias-variance trade-off, high-dimensional datasets and complex modeling procedures can make interpretation difficult in machine learning (Cohen & DeRubeis, 2018; MacEachern & Forkert, 2021; Mooney & Pejaver, 2018). Fortunately, advances in interpretable machine learning (e.g., SHAP method for feature importance (Lundberg & Lee, 2017)) can help to counteract this concern. These techniques allow us to consider many features across categories while identifying which features contribute most to model performance.

Finally, machine learning provides rigorous resampling techniques (e.g., cross-validation) to fit and evaluate models in separate data (James et al., 2013). Models are likely to 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 (Jonathan et al., 2000; Krstajic et al., 2014). 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.

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 (Cornelius, 2020; National Center for Chronic Disease Prevention and Health Promotion (US) Office on Smoking and Health, 2014; Schlam & Baker, 2013). 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 (Jha Prabhat et al., 2013).

Although rates of smoking have declined considerably, approximately 14% of U.S. adults continue to smoke daily or near-daily (Cornelius, 2020); among individuals 12 and older, this constitutes more than 24 million people (Substance Abuse and Mental Health Services Administration, 2023). Additionally, cigarette smoking rates remain much higher in potentially vulnerable populations. This includes people with chronic or severe mental illness (Baker & McCarthy, 2021); Native American and non-Hispanic Black individuals (Cornelius, 2020; Jamal et al., 2015); individuals who are economically and educationally disadvantaged (Cornelius, 2020; Jamal et al., 2015); people with other substance use disorders (Kelly et al., 2012); people in the criminal legal system (Cropsey et al., 2004; Harrison et al., 2020); people experiencing homelessness (Baggett et al., 2013; Soar et al., 2020); people who are insured through Medicaid or uninsured (Cornelius, 2020; Jamal et al., 2015); and individuals who identify as lesbian, gay, or bisexual (Cornelius, 2020; Jamal et al., 2015).

Smoking cessation treatment

Quitting smoking is extremely difficult. Best estimates suggest it takes about 30 attempts to quit successfully; given that as many as 40-50% of former smokers quit on their first serious attempt, there are many individuals for whom it takes abundantly more attempts (2016). Individuals are less likely to achieve sustained abstinence when they have one or more failed quit attempts within the past year (Partos et al., 2013), highlighting the need for individuals to receive effective treatment as early as possible.

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 (Baker & McCarthy, 2021; Cahill et al., 2013; Rigotti et al., 2022; Schlam & Baker, 2013). 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 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 counseling (Cahill et al., 2013; Fiore et al., 2008). Treatment with an FDA-approved medication doubles the rate of treatment success (Baker & McCarthy, 2021). 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 such as varenicline and C-NRT have comparable population-level effectiveness rates, population effectiveness alone cannot guide selection among smoking cessation treatments. Indeed, 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). These facts suggest a critical need for precision mental health approaches in the cigarette smoking domain.

Previous precision mental health & machine learning research

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

There is a reasonable body of evidence identifying "prognostic" factors (Cohen & DeRubeis, 2018) related to cigarette smoking treatment success. This research aims to predict who will or will not be able to quit successfully (Etter et al., 2023; Issabakhsh et al., 2023; Kaufmann et al., 2015; Kaye et al., 2020; Lai et al., 2021; Megan E. Piper, Schlam, et al., 2017). A systematic review identified that features that predict treatment success span many categories: economic, environ-

mental, sociodemographic, psychological, and physical health variables; engagement in treatment; biomarkers; neurocognitive factors; and smoking use, history, and severity characteristics (Bickel et al., 2023).

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" ((Massago et al., 2024), 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 (Coughlin et al., 2020).

This research investigating prognostic factors is critical for understanding mechanisms that underlie treatment success. Progress in this area may allow us to identify processes that improve the chance of quitting successfully, and we may be able to intervene in some of these mechanisms 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 (Bickel et al., 2023). 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 feature that predict 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" factors (Cohen & DeRubeis, 2018), or factors that determine which treatment to prescribe so that we can 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 (Chen et al., 2018)). Chen and colleagues found that individuals of African American ancestry with a specific variant in one gene (cholinergic receptor nicotinic alpha 5 subunit) had statistically significantly higher treatment success for varenicline compared to C-NRT and placebo (Chen et al., 2020). 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 (Chenoweth et al., 2016; Glatard et al., 2017; Lerman et al., 2015; Schnoll et al., 2009; Shahab et al., 2019), though a recent review notes that "real-world evaluations of NMR to personalize treatment for smoking cessation have produced mixed findings so far" ((Siegel et al., 2020), p. 265).

Although this research holds some potential for selecting among treatments, using biologic and genetic factors comes with downsides when considering accessibility and implementation. The necessary technology is improving, but the testing required to collect these data is neither widely available nor accessible (MacEachern & Forkert, 2021). Access to treatment is a known barrier in mental healthcare and a contributing factor driving mental healthcare disparities (Jacobson et al., 2022). 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, affordability is necessary but insufficient to confer accessibility (Siegel et al., 2020). Additionally, there remain methodological challenges associated with establishing NMR cut-off points for slow vs. fast metabolizers, particularly among

vulnerable sub-populations 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; (Siegel et al., 2020)). In contrast to biological or genetic features, other individual difference features may be better positioned for implementation. For example, many features could be measured easily via self-report, which may enable remote assessment for individuals without insurance or access to in-person medical care.

Other research has investigated these types of 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. They did not find that varenicline's effects varied as a function of drinking status (Kaye et al., 2020). 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; (Piper et al., 2016)). They followed up by exploring moderators of treatment main effects and interactions and found that psychiatric history moderated treatment success for some treatments (Megan E. Piper, Schlam, et al., 2017). They also looked at whether various treatment components affected their purported treatment mechanisms to help elucidate how these treatments may be producing their effects (Megan E. Piper, Cook, et al., 2017). These studies took advantage of largely self-report data, making any resulting models more clinically implementable. However, each study examined only a single factor or examined each factor 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 success.

Opportunities for treatment selection

Although there is not yet much research selecting among smoking cessation treatments, 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; Wang et al., 2009; Zheng et al., 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. Although this heterogeneity is typically neglected when selecting among available treatments, precision mental health approaches would instead take advantage of it.

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 (Cahill et al., 2016). 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 neurotrans-

mitter, 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, 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 (nicotine from cigarettes) activity at the receptor. This would be expected to reduce the pharmacological effect of nicotine, likely reducing the behavioral pleasure of smoking (Cahill et al., 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. Research identifying prognostic factors described in the previous section has shown there are *many* factors that predict smoking cessation, and these factors span *many* clinical and behavioral domains (Bickel et al., 2023). It is possible that some of these prognostic factors may act as prescriptive factors as well (i.e., be able to help prescribe/select a treatment), but this has not yet been tested. Moreover, these many factors have not been considered simultaneously in a model as may be needed to unpack complex clinical phenomena like differential treatment success.

For example, some cigarette smokers may have strong cravings with good self-monitoring. These characteristics may make a treatment like C-NRT, which includes 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, an individual with strong craving *without* good self-monitoring may be less likely to succeed with C-NRT because they cannot identify their moments of need for lozenges or gum. For these individuals, C-NRT may not offer benefit over a single NRT like nicotine patch alone despite its higher effectiveness in the broader population (Cahill et al., 2013). Instead, this multi-component treatment may feel overwhelming, reducing adherence and ultimate effectiveness. Some people who smoke may be prone to side effects from a specific treatment, causing them to discontinue treatment prematurely, though they may not have had the same adverse reactions to a different treatment. Environmentally, an individual who lives with other individuals who smoke may benefit from a partial agonist treatment like varenicline because any secondhand smoke would produce less effect.

Any of these factors, 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 guide treatment selection for smoking cessation. They also point to the value of analytic techniques that can incorporate many features simultaneously and consider 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 may be able to reveal unanticipated features or interactions among 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 (Fiore et al., 2008; Tønnesen et al., 1993). Indeed, Gonzales and colleagues found that "abstinence rates are more than threefold lower for

NRTs" during re-treatment compared to initial treatment using the same medication ((Gonzales et al., 2014), p. 391). A pilot trial showed that treatment adherence improves when individuals are given the opportunity to sample various NRT medications pre-quit (Cropsey et al., 2017), and meta-analyses show only smokers who are highly dependent may benefit from 4 mg (vs. 2 mg) nicotine gum (Lindson et al., 2019). These data suggest differential preferences and even differential effectiveness despite a shared pharmacological mechanism of action across NRT medications. The research previously described using the biomarker NMR also supports differential treatment success (Siegel et al., 2020). Additionally, clinical research has demonstrated differential treatment success on an individual basis for other mental health disorders (e.g., antipsychotic medications for schizophrenia (Roussidis et al., 2013); psychosocial interventions for depression (Cohen & DeRubeis, 2018)), suggesting it is worth investigating whether the same is true in smoking cessation.

In summary, many opportunities exist for differential treatment selection among smoking cessation treatments. Pharmacological and behavioral factors may guide selection alone or interaction with one another. The features that previous research has identified that can predict treatment success may interact with specific treatment effects to guide selection, and unanticipated features or interactions may also come to light. Using machine learning analytic techniques allows us to explore these many categories of features simultaneously to capture the many, small effects that may combine to permit treatment selection.

Specific Aims

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. Specifically, we pursued the following aims:

AIM 1: Build a machine learning model to predict cigarette smoking cessation. We built a model that could predict treatment success at 4 weeks post-quit. This model used retrospective data from a previously completed comparative effectiveness trial wherein individuals who smoke were randomized to treatments, richly characterized at baseline, and followed to assess short- and long-term treatment success (Baker et al., 2016).

AIM 2: Evaluate the clinical benefit of using our prediction model for differential treatment selection. We used the prediction model to identify the best treatment for each person and evaluate whether individuals who received their best treatment were more likely to be abstinent in the original trial at 4 weeks, 12 weeks (end-of-treatment), and 6 months post-quit.

Across both AIMS, we used rigorous resampling techniques to ensure that we evaluated our model's capacity to predict smoking cessation outcomes and select treatments in *new* patients. Additionally, we incorporated easy-to-collect self-report data as model inputs, and we employed a statistical algorithm that can reduce assessment requirements. These choices position our model for accessible clinical implementation.

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 on our OSF page and as analysis notebooks on our study website.

We preregistered our analyses to evaluate clinical benefit that relied on significance testing. The preregistration can be found on our OSF page. These analyses also closely followed published

research evaluating treatment selection models (DeRubeis et al., 2014). For our Bayesian hierarchical generalized linear models, we followed guidelines from the tidymodels team (Kuhn, 2022) that we have followed in other published research from our laboratory (Wyant et al., n.d.). 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) (Baker et al., 2016). 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 looking to quit smoking were enrolled in Madison, WI, USA and Milwaukee, WI, USA. Exclusion criteria included contraindicated medical (e.g., severe hypertension) or psychiatric (e.g., severe and persistent mental illness) conditions; 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 one of three medication conditions: varenicline, C-NRT, or nicotine patch. Each medication treatment lasted 12 weeks. 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.

All participants also received 6 sessions of motivational and skill-training counseling per clinical guidelines (Fiore et al., 2008).

Individual difference characteristics

Participants were comprehensively assessed for individual difference 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 1.

Table 1: Individual Differences Characteristics Available for Model Features

Demographic Characteristics	Feature Name	Type	# Items
	Gender	Categorical (unordered)	1
	Age	Numeric	1
	Race	Categorical (unordered)	1
	Marital Status	Categorical (unordered)	1
	Income	Categorical (ordered)	1
	Ethnicity	Categorical (unordered)	1
	Employment	Categorical (unordered)	1

Table 1: Individual Differences (Continued)

Smoking Use/History	Feature Name	Type	# Items
	Baseline Carbon Monoxide	Numeric	1
	Carbon Monoxide Exposure	Categorical (unordered)	1
	Age of 1st Cigarette	Numeric	1
	Age Became Daily Smoker	Numeric	1
	Years Smoking	Numeric	1
	Cigarettes Per Day (Heaviest)	Numeric	1
	Use of Other Tobacco Products	Categorical (unordered)	5
	Number of Previous Quit Attempts	Numeric	1
	Last Recent Quit Attempt	Categorical (ordered)	1
	Longest Quit Attempt	Categorical (ordered)	1
	Previous Quit Methods Used	Categorical (unordered)	6
	Cigarettes Per Day (Current)	Numeric	1
	Motivation to Quit	Categorical (ordered)	1
	Self-Efficacy for Quitting in Next 30 Days	Categorical (ordered)	1
	Confidence to Quit	Categorical (ordered)	1
	Importance to Quit	Categorical (ordered)	1
	DSM5 Tobacco Use Disorder (American Psychiatric Association, 2013)	Categorical (unordered)	13
	Fagerstrom Test of Nicotine Dependence (Heatherton et al., 1991)	Categorical (unordered); Categorical (ordered)	6
	Wisconsin Inventory of Smoking Dependence Motives-37 (Smith et al., 2010)	Categorical (ordered)	37
	Smoke Menthol Cigarettes	Categorical (unordered)	1
	Wisconsin Smoking Withdrawal Scale-2 (Smith et al., 2021)	Categorical (ordered)	38

Table 1: Individual Differences (Continued)

Social & Environmental Characteristics	Feature Name	Type	# Items
	Spouse Smokes	Categorical (unordered)	1
	Live with Another Smoker	Categorical (unordered)	1
	People Close to You Who Smoke	Categorical (unordered)	5
	Ban on Smoking at Home	Categorical (unordered)	1
	Ban on Smoking at Work	Categorical (unordered)	1
	Time Around Other Smokers	Categorical (ordered)	2

Mental Health & Psychological Traits	Feature Name	Type	# Items
	Frequency of Drinking Alcohol	Categorical (ordered)	1
	Quantity of Alcohol	Categorical (ordered)	1
	Binge Drinking	Categorical (ordered)	1
	Short Inventory of Problems-2 (Revised) (Kiluk et al., 2013)	Categorical (ordered)	15
	Life Satisfaction	Categorical (ordered)	1
	Life Enjoyment	Categorical (ordered)	1
	Psychological Disorder Diagnoses	Categorical (unordered)	7
	Positive and Negative Affect Schedule (Crawford & Henry, 2004)	Categorical (ordered)	6
	Snaith-Hamilton Pleasure Scale (Snaith et al., 1995)	Categorical (ordered)	14
	Anxiety Sensitivity Index-3 (Taylor et al., 2007)	Categorical (ordered)	18
	Distress Tolerance Scale (Simons & Gaher, 2005)	Categorical (ordered)	15
	Patient History Questionnaire-9 (Kroenke et al., 2001)	Categorical (ordered)	9

Table 1: Individual Differences (Continued)

Medical & Physical Health	Feature Name	Type	# Items
	Diabetes Diagnosis	Categorical (unordered)	1
	Multidimensional Fatigue Inventory (Smets et al., 1995)	Categorical (ordered)	20
	Berlin Sleep Questionnaire (Netzer et al., 1999)	Categorical (ordered), Categorical (unordered)	3
	Body Mass Index	Numeric	1
	Health-related Quality of Life Scale (Taylor & National Center for Chronic Disease Prevention and Health Promotion (U.S.). Division of Adult and Community Health, 2000)	Categorical (ordered), Numeric	4
	Healthy Days Symptoms Module (Moriarty, 1996)	Numeric	5

Miscellaneous Features	Feature Name	Type	# Items
	Treatment	Categorical (unordered)	1
	Incarcerated	Categorical (unordered)	1
	Works Third Shift	Categorical (unordered)	1

Treatment success outcome

Throughout study participation, treatment success was assessed via biologically confirmed, 7-day point-prevalence abstinence. Point-prevalence abstinence assesses for any smoking (here, in the previous 7 days) and yields a single, dichotomous outcome of "abstinent" or "smoking." 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 any 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."

Assessments of treatment success were used in two separate ways across our two AIMS.

In AIM 1, treatment success served as the *prediction outcome for our models*. For our primary analyses, we built prediction models to predict treatment success at 4 weeks post-quit (i.e., predicting if individuals were labeled "abstinent" or "smoking" at 4 weeks). We also conducted supplemental analyses where we built models predicting treatment success at 12 and 26 weeks post-quit (see Supplement).

In AIM 2, treatment success served as the outcome for *clinical benefit analyses*. We evaluated whether using our treatment selection model yielded higher treatment success (i.e., higher rates of abstinence) at 4, 12, and 26 weeks post-quit.

AIM 1 analytic strategy

Model configurations

Machine learning seeks to optimize the bias-variance trade-off: too many features can yield high-variance models that do not generalize well to new data; too few features can yield biased models that miss important predictive relationships. It is difficult to determine *a priori* the optimal trade-off between bias and variance. Consequently, we considered many model configurations that differed across characteristics expected to affect bias and variance. These characteristics include feature engineering steps, dimensionality reduction approaches, different size feature sets, and hyperparameter values. The model configuration that most closely captures the optimal bias-variance trade-off will perform best in new data.

Statistical algorithm. We 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 features). This permits considering many interactions that each account for a small portion of variance, which may be necessary to capture the complexity of clinical phenomena like treatment success.

Second, GLMNet performs a degree of dimensionality reduction that may be helpful for implementation. It penalizes model complexity via two hyperparameters (alpha and lambda) such that features are retained in the model only when their contribution to prediction outweighs the cost of adding another parameter to the model. A model using this algorithm may ultimately 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. As an additive linear model, the predictions come from the sum of the retained features multiplied by their parameter estimates. This makes it easier to understand how the model is making its predictions, as compared to "black box" statistical algorithms that offer no insight as to what is happening under the hood (Kuhn & Johnson, 2018). The dimensionality reduction that GLMNet performs also increases its interpretability by reducing the high-dimensional feature space to a more manageable set of features. We aimed to prioritize interpretability to avoid a potential barrier to using machine learning approaches for clinical and public health goals (Cohen & DeRubeis, 2018; MacEachern & Forkert, 2021; Mooney & Pejaver, 2018).

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.

Feature engineering. Feature engineering is the process of converting raw data into meaningful features that improve model effectiveness (Kuhn & Johnson, 2019). 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 data, mode imputation for categorical [ordered and unordered] data); 2) removing zero-variance features; 3) using a Yeo-Johnson transformation on all numeric features to normalize distributions; 4) one-hot-coding unordered categorical data; and 5) standardizing all features to have a mean of 0 and standard deviation of 1 (as a requirement of the GLMNet algorithm). Medians/modes for missing data imputation, identification of zero variance features, and means

and standard deviations for normalization and standardization were derived from held-in (training) data and applied to held-out (validation and test) data (see Cross-validation section below).

Treatment was also one-hot-coded such that there were three features, one corresponding to each treatment (varenicline, C-NRT, nicotine patch). The three treatment features were allowed to interact with all other features to permit differential prediction.

Feature sets differed in two ways across model configurations. First, ordered categorical data (e.g., Likert scale items on self-report measures) could be ordinal scored (i.e., treated as numeric features) or one-hot-coded (i.e., treated as unordered categorical features). Ordinal scoring may decrease variance as it would create fewer features (e.g., single numeric feature ranging from 1 - 7) compared to one-hot-coding (e.g., 7 features for an item with 7 response levels). However, one-hot-coding allows for non-linear relationships and thus more flexibility in predictive patterns, potentially decreasing bias. Second, we included *either* items (i.e., individual items within a self-report measure) or scale scores (i.e., total scale and sub-scale scores derived from items in a self-report measure). Feature sets with items had far more features and allowed for unequal weighting of items within a single scale, decreasing bias. Feature sets with scale scores had fewer features and required that all items that comprised the scale score were weighted equally (e.g., sum score), decreasing variance.

Dimensionality reduction. Dimensionality reduction approaches offer additional ways to produce lower-variance solutions that may generalize better to new data. We used several data-driven methods for dimensionality reduction such as: removing near-zero variance features, removing highly correlated features (via GLMNet penalization), and including lower-dimensional feature sets. We also used several non-data-driven approaches for dimensionality reduction. First, we removed variables that conflicted with our ultimate implementation goals (e.g., variables whose assessment required blood work or lab tests). This reduced the overall number of features in the model. Second, we removed variables that lacked face validity for predicting abstinence (e.g., detailed questions about snoring used to diagnose sleep apnea; retained sleep apnea diagnosis) or overlapped conceptually with other features (e.g., including body mass index instead of height and weight separately).

Model fitting, selection, & evaluation

Cross-validation. Cross-validation allows for rigorous consideration of many model configurations and prioritizes performance in new data not used for model training (Jonathan et al., 2000). Specifically, we used nested cross-validation for model training, selection, and evaluation (Krstajic et al., 2014).

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 (held-in data), select the best models (held-out validation sets), and evaluate those best models (held-out test sets). 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 (Jonathan et al., 2000; Krstajic et al., 2014).

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 model performance across the 10 *validation sets*. Final performance evaluation of those best model configurations used median performance across the 30 *test sets*.

Metrics. Our primary performance metric for model selection and evaluation was area under the Receiver Operating Characteristic Curve (auROC) (Kuhn & Johnson, 2018; Youngstrom, 2014). In simplest terms, auROC measures how well our model discriminates between positive

(abstinent) and negative (smoking) cases. More specifically, auROC indexes the probability that the model will predict a higher score for a randomly selected positive case relative to a randomly selected negative case. It ranges from 0.5 (chance prediction) to 1.0 (perfect prediction). This metric was selected because it 1) combines sensitivity and specificity, which are both important characteristics for clinical implementation; 2) is unaffected by class imbalance; and 3) does not automatically use a threshold, which is not needed when working with predicted probabilities rather than class predictions.

Although we can only optimize on one metric (i.e., auROC), we can define other relevant criteria as "satisficing metrics" ((Ng, 2018), p. 22). We can require that our models meet a specific value for the satisficing metric and then optimize among those models. Specifically, because our goal was treatment selection, it was important that our models retain interaction features. Although retaining interaction features does not guarantee that our models will be able to select among treatments, *not* retaining any interactions *prevents* treatment selection.

We defined a satisficing metric of 50 or more treatment interaction features retained. This value was derived from the inner loop of nested cross-validation. As with auROC, we calculated the median number of treatment interaction features retained for each model configuration (i.e., averaged across the 10 validation sets). We arrived at a criterion of 50 interaction terms retained because: 1) this represented approximately the 50th percentile in the distribution of median interaction terms retained across model configurations; and 2) using this threshold still ensured there were sufficient model configurations among which to choose in each outer fold.

Thus, we selected as our best models the models which 1) had the highest median auROC across folds, while 2) retaining a minimum median 50 treatment interaction features. 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.

Bayesian analysis of model performance. We used a Bayesian hierarchical generalized linear model to estimate the posterior probability distributions and 95% Bayesian credible intervals (CIs) for auROC for the best models. The posterior probability updates a prior probability (derived from a distribution) using new evidence provided by our data. A posterior probability distribution around a given parameter (e.g., median auROC) is the estimated distribution of these probabilities that allow us to make statements about the likelihood of observing our results.

We estimated the posterior probability distribution around model performance following recommendations from the tidymodels team (Kuhn, 2022). We regressed the auROCs (logit transformed) from the 30 test sets on 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 model's auROCs. If 95% Bayesian CIs do not include 0.5 (chance performance), we can conclude that the model performs better than chance.

Bayesian analyses were accomplished using the tidyposterior (Kuhn, 2022) and rstanarm (Goodrich et al., 2023) packages in R. Following recommendations from the rstanarm team and others (Gabry & Goodrich, 2023; RStudio Team, 2020), 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.

Model interpretation

Interpretability is important for several reasons when using machine learning approaches for clinical applications (Cohen & DeRubeis, 2018; MacEachern & Forkert, 2021; Mooney & Pejaver, 2018). First, understanding how the model is making predictions, including identifying important features, can help advance theory. For example, determining factors that predict smoking cessation may

help to guide novel treatment development by identifying maintaining mechanisms or predictors of success. Second, interpretability may help support implementation, as patients or clinicians may be more likely to use a model that they understand.

We pursued model interpretation in two ways. First, as described above, the GLMNet statistical algorithm offers several advantages for interpretability. It is a linear model that outputs parameter estimates for each feature. It also performs regularization by shrinking parameter estimates and/or removing unnecessary or highly correlated features from the model entirely. Features are retained in the model only if their contribution to performance outweighs the cost of having an additional parameter in the model. Consequently, we can review the retained features and their parameter estimates as a metric of feature importance.

Second, we computed Shapley Values (Lundberg & Lee, 2017) 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 using the final, best selected model configuration. Leave-one-out cross-validation works identically to k-fold cross-validation described above. The value of k is set to N (sample size) such that each test set consists of a single held-out participant. We fit $N = 1086$ models, where each participant served as the test set once for a model fit with the other 1085 participants. This method 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 (see "Select final model configuration" below).

We used the DALEX and DALEXtra packages (Biecek, 2018) 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 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.

The additivity property of Shapley values also allowed us to create two feature categories, Main Effects and Interactions. Therefore, in addition to identifying the most important individual features, we were able to assess the relative contribution of treatment interaction features and main effect (non-interaction) features.

AIM 2 analytic strategy

We followed methods described and used previously for evaluating the potential clinical benefit of treatment selection models (DeRubeis, 2019; DeRubeis et al., 2014). All analyses for Aim 2 were preregistered. The preregistration can be found on our OSF page (https://osf.io/qad4n/).

Select final model configuration

Nested cross-validation evaluates how well a model selected with cross-validation will perform on a held-out test set (Krstajic et al., 2014). To get a single, final model, we replicated our inner loop resampling (1 repeat of 10-fold cross-validation) on the full dataset. The best model configuration was selected using median auROC across the 10 held-out folds.

A final model was fit on the full dataset using this best selected model configuration. This model was used to obtain parameter estimates to aid with interpretation, and it would serve as the model disseminated for ultimate implementation.

Identify model-predicted best treatment

As we did for feature importance with Shapley values, we conducted leave-one-out cross-validation 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 predicted probabilities for each participant by substituting each treatment into the model inputs. This produces 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 evaluated the clinical benefit of our model-selected treatment by comparing 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 therefore 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, but 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 (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

All 1086 participants who were randomized to treatment in the comparative effectiveness trial (Baker et al., 2016) were included in our analysis sample.

Demographic characteristics of this sample appear in Table 2. Our sample was 52.12% (N = 566) female. We had good representation of White (67.03%, N = 728) and Black (28.45%, N = 309) individuals but poor representation among Asian (0.28%, N = 3), Multiracial (2.03%, N = 22), Native American/Alaska Native (0.55%, N = 6), and Other individuals (1.66%, N = 18). This was also a primarily non-Hispanic sample (97.51%, N = 1059). There was wide variety with respect to marital status, employment status, and income, including almost one third of the sample reporting annual income lower than $20,000.

Table 2: Demographic Characteristics

Characteristic	N (%)	Mean (SD)
Age		48.13 (11.6)
Gender		
Female	566 (52.12%)	
Male	520 (47.88%)	
Race		
Asian	3 (0.28%)	
Black/African American	309 (28.45%)	
Multiracial	22 (2.03%)	
Native American/Alaska Native	6 (0.55%)	
Other	18 (1.66%)	
White	728 (67.03%)	
Ethnicity		
Hispanic or Latino/a	27 (2.49%)	
Non-Hispanic	1059 (97.51%)	
Marital Status		
Divorced	224 (20.63%)	
Living with a domestic partner	87 (8.01%)	
Married	384 (35.36%)	
Never married	299 (27.53%)	
Separated	53 (4.88%)	
Widowed	34 (3.13%)	
Did not respond	5 (0.46%)	
Employment		
Employed (full-time)	474 (43.65%)	
Employed (part-time)	283 (26.06%)	
Unemployed	329 (30.29%)	

Table 2: Demographic Characteristics (Continued)

Characteristic	N (%)	Mean (SD)
Income		
< $10,000	210 (19.34%)	
$10,000 - $19,999	135 (12.43%)	
$20,000 - $24,999	83 (7.64%)	
$25,000 - $34,999	129 (11.88%)	
$35,000 - $49,999	149 (13.72%)	
$50,000 - $74,999	160 (14.73%)	
$75,000+	167 (15.38%)	
Did not respond	53 (4.88%)	

Smoking-related characteristics of this sample appear in Table 3. On average, individuals had been smoking almost 30 years and were currently smoking 17 cigarettes per day. Over three quarters of the sample reported smoking their first cigarette of the day within 30 minutes of waking up.

Table 3: Smoking Use & History Characteristics

Characteristic	Mean (SD)	N (%)
Age of first cigarette	14.6 (3.42)	
Age became daily smoker	17.56 (4.62)	
Years smoking	28.65 (12.03)	
Cigarettes per day (heaviest)	22.74 (10.67)	
Number of previous quit attempts	3.91 (6.03)	
Cigarettes per day (current)	17.03 (8.31)	
Number of DSM5 tobacco use disorder symptoms	4.57 (2.17)	
WISDM37 Total Score	3.99 (1.17)	
WSWS Total Score	48.82 (19.91)	
Time to first cigarette upon waking		
After 60 minutes		82 (7.55%)
31 - 60 minutes		157 (14.46%)
6 - 30 minutes		477 (43.92%)
Within 5 minutes		366 (33.7%)
Did not respond		4 (0.37%)

AIM 1 results: Prediction models

Model performance

We selected the best model configurations using auROCs from the *validation sets* (among models that met our satisficing metric of a median of at least 50 retained treatment interaction features across inner folds). We evaluated these best model configurations using *test set* performance to remove optimization bias present in performance metrics from validation sets (Krstajic et al., 2014). The median auROC across the 30 test sets for the 4-week model was 0.695 (IQR = 0.667 - 0.718, range = 0.592 - 0.788). This value is comparable to model performance from extant literature predicting smoking cessation using machine learning (e.g., auROC = 0.660 (Lai et al., 2021)). Figure 1, Panel A shows the ROC curve for held-out test set performance (concatenated across 30 held-out folds).

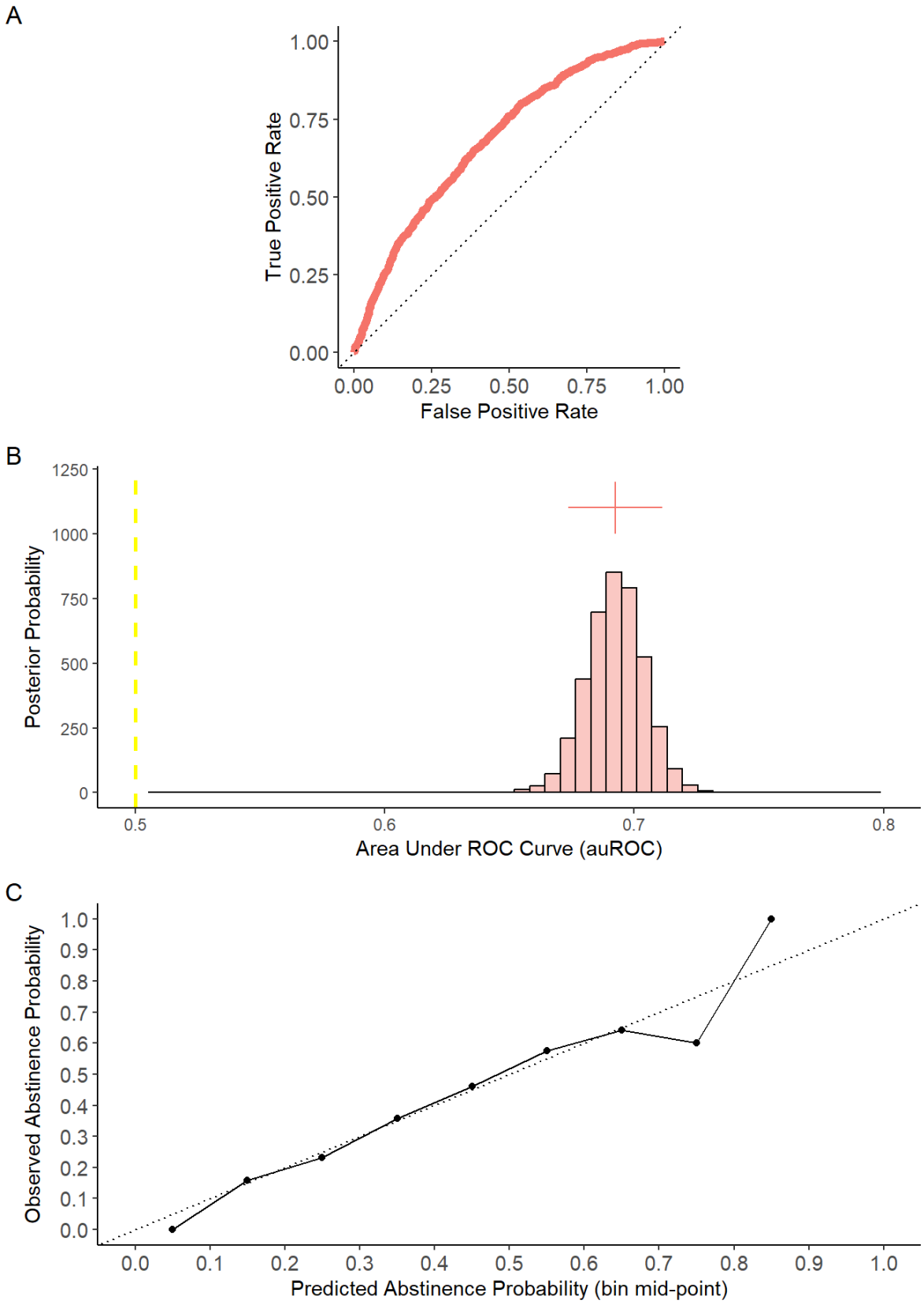


Figure 1: Model performance for prediction model. A) ROC curve plotted across all values of sensitivity (true positive rate) and specificity (1 - false positive rate). Dotted line indicates chance performance. B) Posterior probability distribution for the median auROC in test sets. Histogram represents posterior probability distribution. Horizontal line displays 95% Bayesian credible interval. C) Model calibration between predicted probabilities and observed values. Predicted probabil-

We used the 30 test set auROCs to estimate the posterior probability distribution for the auROC of these models. The median auROC from the posterior distribution was 0.693. This value represents our best estimate for the magnitude of the auROC parameter. The 95% Bayesian CI for the auROC was relatively narrow [0.674 - 0.711] and did not contain 0.5 (chance performance), suggesting this model has predictive signal. Figure 1, Panel B displays the posterior probability distribution for the auROC.

Model calibration

In Figure 1, panel C, we display our model's calibration. Predicted lapse probabilities are binned (bin width = 10%) and plotted against the observed probability of abstinence for observations in that bin. If probabilities were perfectly calibrated, all bin means would fall on the dotted line (e.g., bin from 0 - 10 with an observed mean probability of 0.05, bin from 10 - 20 with an observed mean probability of 0.15). This figure plots the probabilities from our held-out predictions (made with leave-one-out cross-validation) using the final selected model configuration for each individual for their original trial-assigned treatment against the observed trial abstinence rates. Probabilities are well calibrated and ordinal in their relationship with the true probability of abstinence. Given this, these probabilities can provide precise predictions of treatment success that can be used for treatment selection.

Model interpretation

Retained features. Our final model fit with the full dataset retained retained 155 features. Tables with all retained features as well as their precise parameter estimates appear in the Supplement.

Of the retained features, 74 were treatment interaction features. Retained treatment interaction features appear in Table 4, grouped by feature category from Table 1. The effect direction indicates whether increasing values of that feature (or coded positive for that level for one-hot-coded categorical features) increased or decreased the probability of treatment success when using that specific treatment (vs. the other two treatments).

Table 4: Retained Interaction Features

Category	Feature	Effect Direction (Varenicline)	Effect Direction (C-NRT)	Effect Direction (Patch)
Demographic				
	Divorced		-	
	Greater income	+		
	Has never been married			+
	Identifies as Black or African American		-	
	Identifies as White		+	+
	Identifies as female			-
	Identifies as male			+
	Married		+	
Medical				
	Berlin: Feels tired, fatigued, or not up to par	-		
	Berlin: No sleep apnea diagnosis		+	
	HDSM: More days in past month feeling worried, tense, or anxious		-	
	MFI: Does not feel it takes a lot of effort to concentrate		+	
	MFI: Does not have a lot of plans			+
	MFI: Does not think they do a lot in a day	+		
Psychological				
	ASI-3: Worries they are going		-	
Smoking/Environment/ History	[redacted]	+	+	+

To perform treatment selection, only interactive features need to be assessed, as features that increase or decrease probabilities equally across all three treatments do not help with differential prediction. Consequently, implementing this model for treatment selection would require assessing only 52 unique items (e.g., multiple dummy variables are from a single item, the same feature interacts with more than one treatment).

Feature importance via Shapley values. Global feature importance (mean |Shapley value|) from our model appear in Figure 2, both overall (Panel A) and for treatment interaction features specifically (Panel B). Shapley values describe the relative importance of these individual features for making predictions. Six of the top 25 most globally important features were treatment interactions.

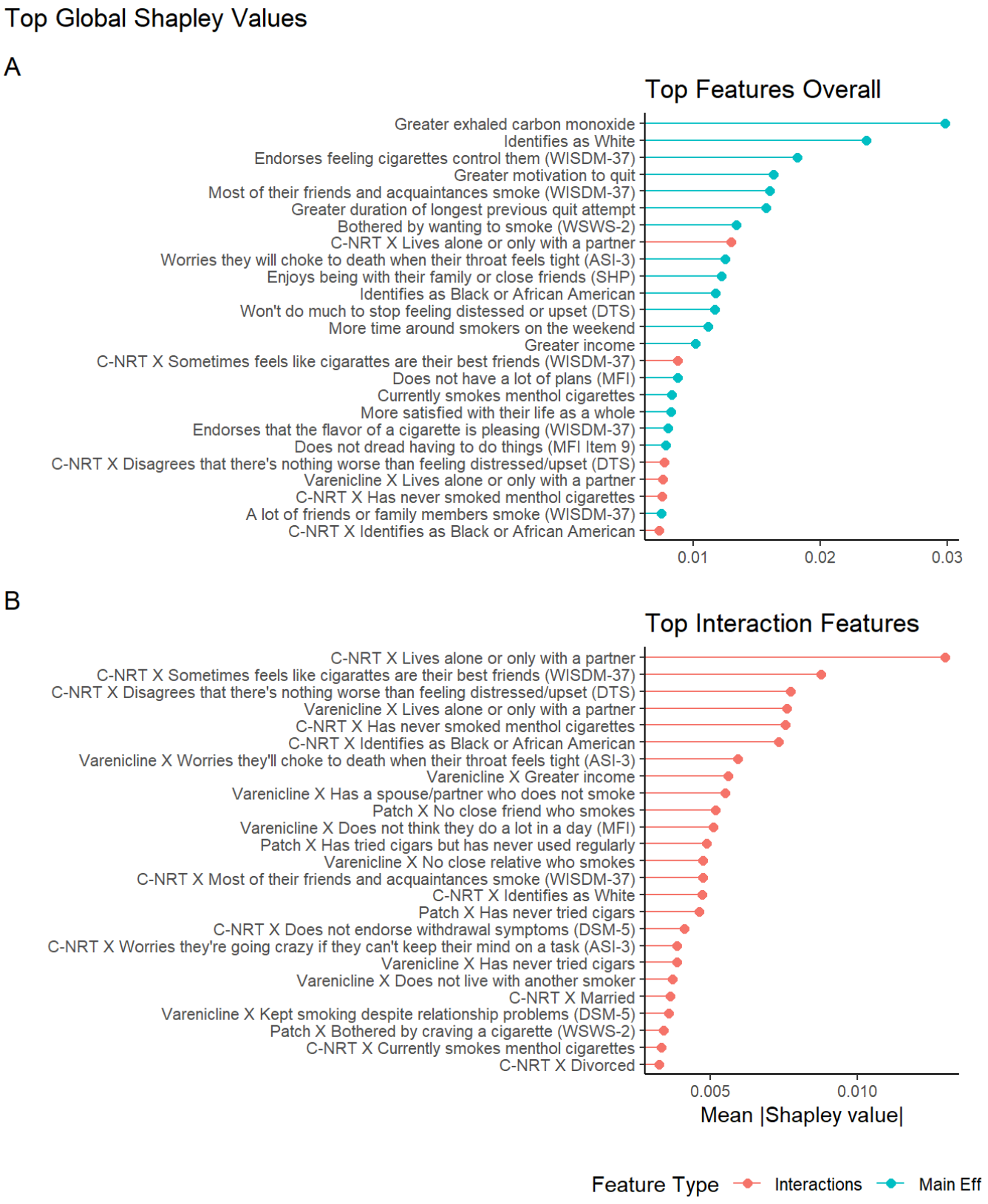


Figure 2: Global feature importance via Shapley values. Bar represents magnitude of global feature importance, which is calculated from the mean of the absolute value across all observations for that feature. A) Global feature importance for top 25 features overall. B) Global feature importance for top 25 treatment interaction features.

Global feature importance grouped by feature type (main effect or interaction) appear in Figure 3. Main effect features were relatively more important than treatment interaction features for predicting overall treatment success.

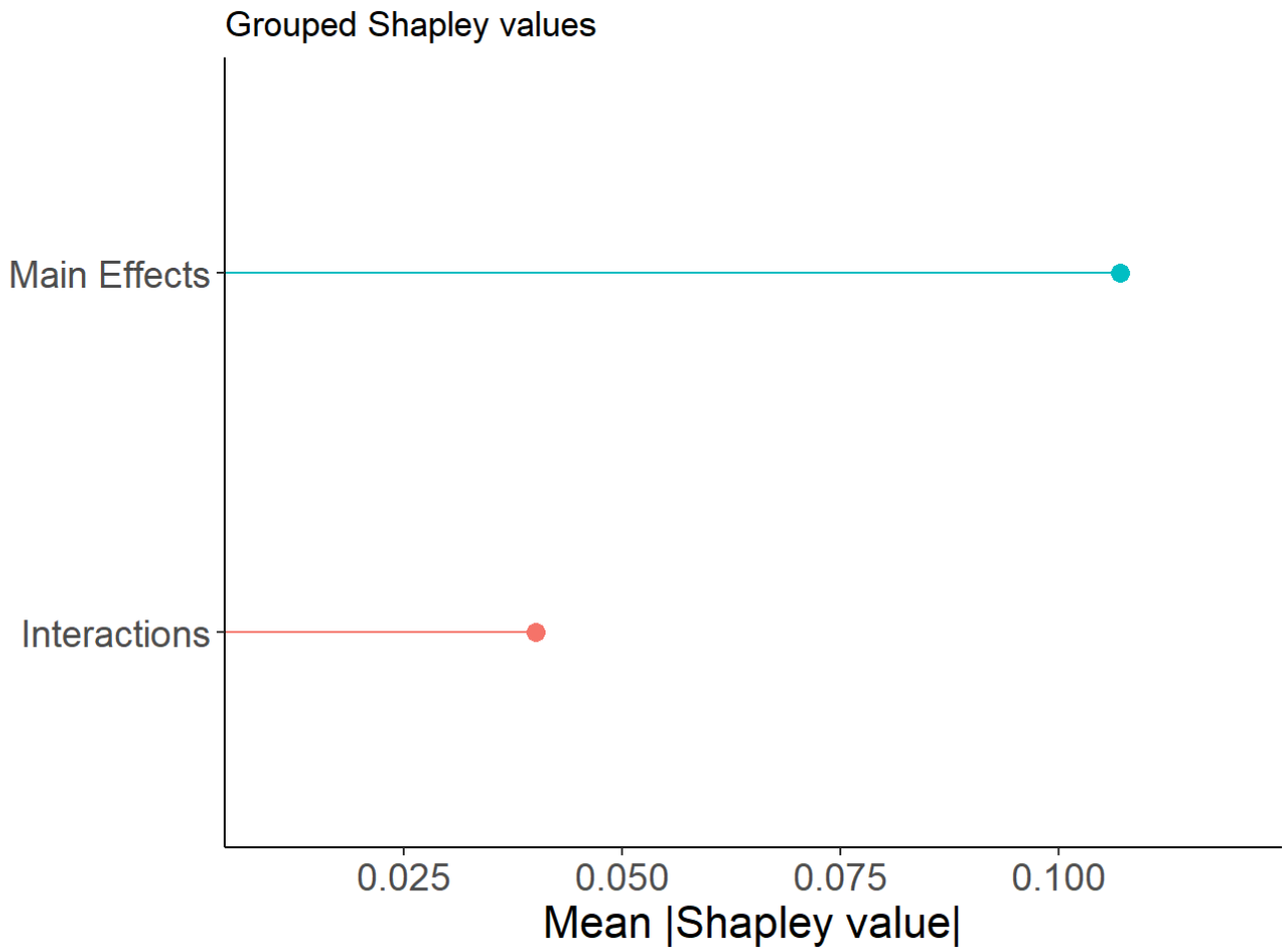


Figure 3: Global feature importance via Shapley values grouped by feature type (main effect or interaction feature)

AIM 2 results: Clinical benefit

There was a significant fixed effect of treatment matching on abstinence (odds ratio [OR] = 1.851, $z = 2.862$, $p = 0.004$). Individuals who received their model-predicted best treatment were more likely to be abstinent than individuals who did not. There was also a significant fixed effect of time (OR = 0.210, $z = -10.053$, $p < 0.001$) such that the probability of abstinence declined over time.

There was not a significant interaction between treatment matching and time ($p = 0.697$). However, calculating and interpreting interactions in logistic models is not straightforward because significance can differ based on the link function used (Collins, 2018; Karaca-Mandic et al., 2012). Consequently, we conducted simple effects analyses of the effect of treatment matching at each time point.[1] This allowed us to characterize our results more fully and to understand our effects in their original probability terms.

[1]We followed our preregistered analysis code to fit these simple effect models; however, we preregistered that we would conduct them only if the interaction effect was significant.

There was a significant fixed effect of treatment matching on abstinence at 4 weeks (OR = 1.382, $z = 2.452$, $p = 0.014$) such that individuals who received their model-predicted best treatment were more likely to be abstinent. The effect of treatment matching was no longer significant at 12 weeks ($p = 0.232$) or at the 26-week follow-up assessment ($p = 0.943$). Figure 4 shows the mean abstinence rate by treatment matching at each time point.

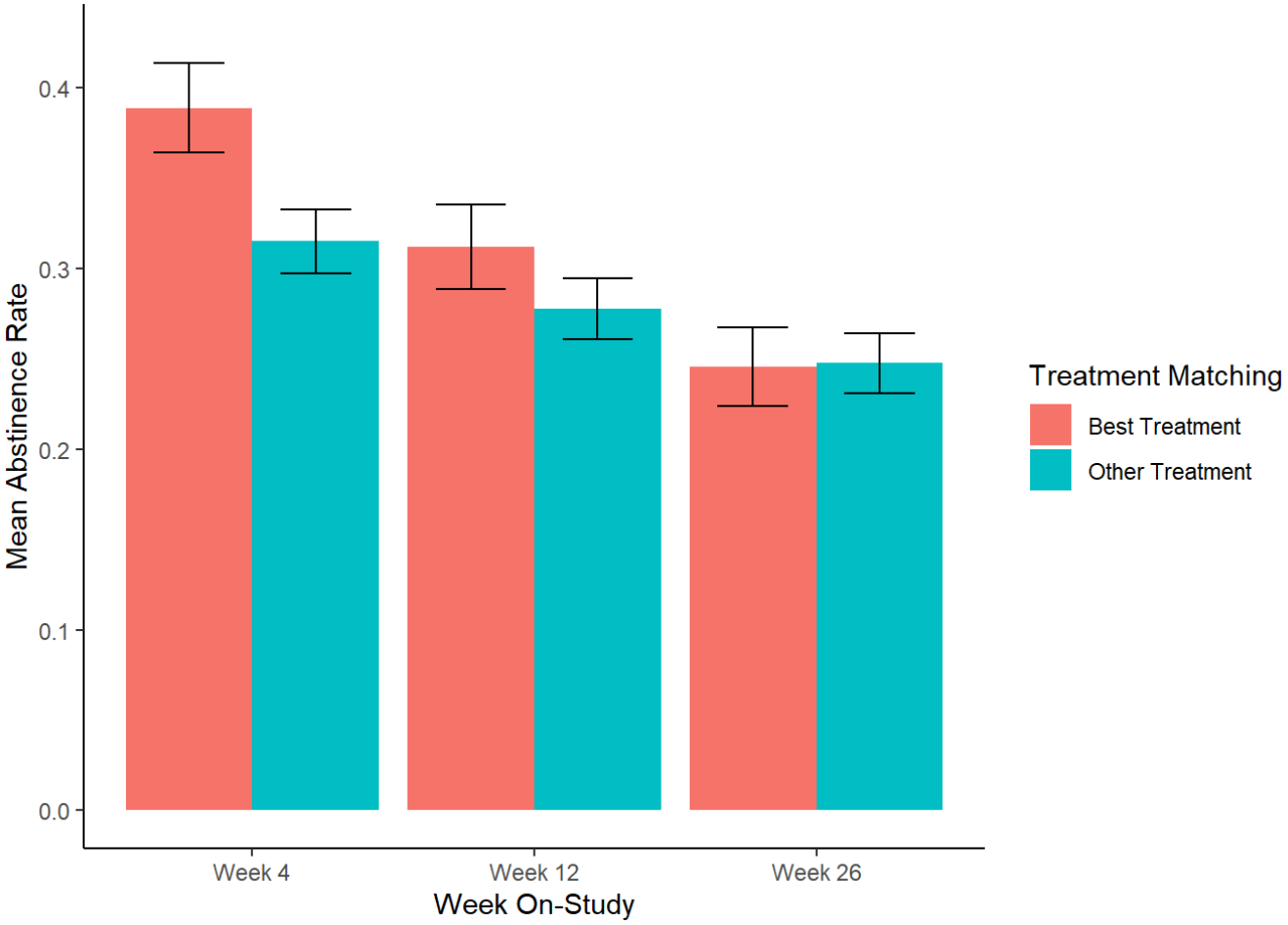


Figure 4: Benefit of treatment matching. Bars represent mean observed abstinence (from original trial) for individuals who did and did not receive their model-predicted best treatment, over time. Error bars indicate standard errors.

Discussion

In this project, we have produced a treatment selection model that can offer immediate benefit to individuals looking to quit smoking using several first-line medications. We can identify the best treatment for a specific individual at the moment in time that they are looking to quit. This treatment selection model can accomplish the goals of the precision mental health paradigm: the right treatment, for the right person, at the right time (Kaiser, 2015).

Benefit of treatment selection

Individuals who received their model-predicted best treatment in the original trial had a mean abstinence rate of 38.9% at 4 weeks. In comparison, the mean abstinence rate at 4 weeks among

people who did not receive their best treatment was 31.5%. Although this difference may seem somewhat small numerically, this represents a vast improvement. Using our model yields another 7.4% increase - almost another quarter beyond what our *best* treatments offer - simply by allocating them to the right person.

We feel confident in this effect for several reasons. First, we made predictions for each individual to identify their best treatment while they were held-out from model fitting. Thus, our treatment selection process matches how this model will be used in clinical practice - to make predictions and select a treatment for new patients. Second, our model is capturing predictive signal, as supported by the Bayesian CI around our model's auROC, adding to our trust in the model's prediction outputs. Third, these predictions were well-calibrated such that we can trust their ordinal ranking. This is critical because what is used in our treatment selection process is the relative *order* (i.e., rank) of predicted probabilities for each person rather than the values themselves.

Moreover, we can achieve this benefit using an assessment that is both low-burden and accessible. Implementing this treatment selection model would require assessing approximately 50 multiple choice and yes/no questions. These types of questions take 10-15 seconds each to answer on average (Lenzner et al., 2010), meaning a 52-item assessment would be expected to take approximately 11-12 minutes. Additionally, because all items are self-report questions, this assessment can be completed remotely (e.g., administered online). Consequently, it can be made available to people who are un- or under-insured or who do not have access to in-person medical care. This assessment tool is particularly valuable in this context because two treatments in the model (C-NRT, nicotine patch) are widely available over-the-counter, offering scalable implementation when healthcare access is limited.

This focus on accessibility is particularly important given disparities in mental healthcare. Access to treatment is a known barrier in mental healthcare and a contributing factor driving healthcare disparities (Jacobson et al., 2022). Cigarette smoking rates remain higher in many marginalized populations (Baggett et al., 2013; Baker & McCarthy, 2021; Cornelius, 2020; Cropsey et al., 2004; Harrison et al., 2020; Jamal et al., 2015; Kelly et al., 2012; Soar et al., 2020). Precision mental health approaches must aim to mitigate rather than exacerbate health disparities in our treatment pipeline; prioritizing accessibility in implementation is a critical first step towards this goal (MacEachern & Forkert, 2021).

Long-term outcomes

Challenges of predicting distal treatment success

Alongside these exciting findings, this study also demonstrates that it is difficult to predict complex outcomes like treatment success using distal predictors. Our best estimate of median auROC in held-out data (from the posterior probability distribution) was 0.693, and our Bayesian CI around the model's auROC indicated that our model is capturing predictive signal. An auROC of 0.693 indicates that our model correctly assigns a higher probability to a positive (abstinent) case than a negative (smoking) case 69% of the time. An auROC of 0.7 to 0.8 is considered "acceptable" (-Mandrekar, 2010). Our model was right on the cusp of this range.

Nevertheless, there is room for improvement in our model's predictive performance. Model performance also degraded when we built models predicting treatment success at 12 weeks (median auROC: 0.665) and 26 weeks (median auROC: 0.629; see Supplement for additional prediction models). Our model fitting, selection, and evaluation process was identical for these models, suggesting that the greater distance between our baseline predictors and abstinence outcomes was the cause of the degraded performance.

It is important to note, however, that our primary goal was *not* overall prediction of treatment success. Capturing some signal with our model was necessary to yield credible predictions and establish the relevance of features, and so our model's modest predictive performance may have some negative impact on our ability to select treatments. However, overall prediction was insufficient for our purpose. A model that predicted treatment success perfectly (i.e., auROC of 1.0) but included no interactions would be useless for treatment selection. Indeed, this fact motivated our use of a satisficing metric for a minimum number of interaction terms retained in our models, even though it is possible that we sacrificed some level of predictive power by imposing that additional criterion.

Improving long-term benefit of treatment selection

Our treatment selection model offers clinical benefit; however, this benefit is short-lived. There is no longer statistically significant benefit of treatment matching at 12 weeks, though there is a numeric difference (31.2% abstinence vs. 27.8% abstinence), and there is no statistical or numeric difference at all by 6 months.

Initial treatment success is critical, especially for cigarette smoking where even reducing smoking or quitting for some period of time can improve health outcomes and life expectancy (Jha Prabhat et al., 2013). Additionally, smoking early in a quit attempt can have strong negative consequences: decreased self-efficacy, reduced treatment adherence, and premature treatment cessation (Schlam & Baker, 2013). Findings such as these highlight that selecting a treatment that increases treatment success in early recovery (i.e., at 4 weeks) is necessary - though not sufficient - for long-term success.

Nevertheless, we were of course hopeful that treatment matching benefits would endure. However, it is perhaps unsurprising that they do not. We believe there are several possibilities for the lack of benefit at later assessment points.

First, it is possible that we could have improved prediction overall and treatment selection if we incorporated biological markers or genetic features given extant literature suggesting their value (Chen et al., 2018). For example, Chen and colleagues found that varenicline (40% abstinence) was more effective than C-NRT or placebo (both 0% abstinence) for individuals of African American ancestry with a specific genotype at 12 weeks (Chen et al., 2020), demonstrating selective benefit at a later time point than the current study. However, caution is warranted in interpreting this effect. The selective benefit came for a sub-population identified by a single genotype from a single candidate gene with a specific ancestry. This subpopulation comprises 13% (N = 35) of individuals of African American ancestry in this study, and only 4% of the overall sample. There was also no longer benefit at the 6-month follow-up, suggesting this genotype may not inform long-term treatment success. Additionally, their study offered no treatment selection guidance for individuals of African American ancestry with different genotypes or for individuals of European ancestry, limiting its clinical utility. Finally, the potential improvement that could come from including biological or genetic features carries an associated cost to implementation given the relative inaccessibility of genetic and biological testing (MacEachern & Forkert, 2021). For these reasons, incorporating these features may not be the best solution for improving the long-term success of treatment selection.

A second possibility is that we failed to include non-biological/non-genetic features that are critical for predicting treatment success later on. Features that predict at 4 weeks may be qualitatively different than those that predict further out, even if they come from the same domains as our current features (e.g., smoking use and history, mental health). This would mean that we would need different inputs for models to predict short-term and long-term treatment success. However, our data come from a large comparative effectiveness trial conducted by a nationally recognized center and designed by foremost experts in the field. The baseline assessment was quite

comprehensive and was based on domain expertise and decades of research. Thus, it seems unlikely that we could be missing enough important features to bridge the gap in our benefit of treatment selection.

Third, it may be that the same features predict treatment success across time. However, because these characteristics can change dynamically *within an individual*, what was the right treatment for them based on their pre-quit characteristics is no longer the right treatment by 12 weeks, 6 months, or beyond. Many of the features used for prediction in this model were based on a single assessment of *current* states - withdrawal, dependence, confidence/motivation to quit, time around other smokers, distress tolerance, depression symptoms, among others. Even features that feel more "static" like employment or marital status can be subject to change. Consider the example of the feature that indicates that someone lives alone or only with their partner. This same feature increases treatment success using C-NRT and decreases treatment success using varenicline in our model (see Table 4). What if this was an important factor that led to C-NRT being selected as an individual's best treatment - and then they move?

Examples like these are the rule rather than the exception when it comes to chronic diseases like substance use disorders. Tobacco and other substance use disorders are dynamic and relapsing; both risk for use and the factors driving that risk fluctuate over time (Brandon et al., 2007). This change over time has been identified as a key barrier that we must overcome to succeed with precision mental health goals in addiction: "an unspoken assumption underlying much research in this area has been that the purported mechanism of a given addiction treatment is static over time… research on this topic has not typically paid appropriate attention to the dynamic nature of addictive behavior and the complexity of the relapse process" (Oliver & McClernon, 2017).

Thus, though important, initial treatment selection is insufficient for chronic, relapsing disorders. Conditions like these require long-term, continuing care - a fact supported by the declining abstinence rates over time (i.e., main effect of time) in this sample, which is consistent with decades of research (Baker & McCarthy, 2021). Future precision mental health research must take into account both this need for long-term care and the reality of the dynamic nature of tobacco and other substance use disorders.

This will require ongoing assessment of key features that predict treatment success. Such monitoring is now feasible given developments in personal sensing (i.e., in situ data collection via sensors embedded in individuals' daily lives) (Bae et al., 2018; Chih et al., 2014; Epstein et al., 2020; Soyster et al., 2022). Sensing via ecological momentary assessment is well-positioned to capture the self-report items used as features in this treatment selection model, though answering a 50-item survey frequently is likely not feasible. However, it may be that when features are measured much more proximally, they have greater predictive power, and fewer features are needed for successful treatment selection.

In addition to dynamic monitoring, we must be able to select treatments and supports adaptively over time. Adaptive treatment selection may include adjustments to medications such as changing doses or changing medications entirely. It also may include offering other traditional treatments (e.g., psychosocial counseling). Given the chronic, relapsing nature of tobacco and other substance use disorders, these traditional treatments may not be sufficient for long-term care. Alternative treatments and supports via web-based interventions and mobile health apps may offer platforms for sustainable, scalable, ongoing support. Rather than changing treatments entirely or starting a new treatment package (e.g., 12 weeks of counseling), we can recommend specific modules and individual tools that map onto currently important features affecting treatment success. This mapping between risk factors and supports is likely to be quite complex and will require considerable future research. But if we hope to advance precision mental health for smoking, addiction,

and even mental health broadly, we must consider the dynamic nature of risk and recovery inherent in these conditions.

Interpreting our treatment selection model

Several recent reviews have noted that the low interpretability of "black box" machine learning models may impede their utility for clinical and public health goals (Cohen & DeRubeis, 2018; MacEachern & Forkert, 2021; Mooney & Pejaver, 2018). Consequently, we aimed to make our model as interpretable as possible. We used a relatively interpretable statistical algorithm, GLMNet, and we calculated Shapley values to understand relative feature importance among the features in our model.

Prescriptive factors for treatment selection

Using these methods for interpretable machine learning, we were able to identify *prescriptive* features that predict differential treatment success (i.e., features that interact with treatment to help us select among treatments). Our best selected model configuration retained 74 interaction terms that spanned 52 unique items. Each feature's associated global Shapley value, which indicates overall magnitude of feature importance, was relatively small. This finding supports what has long been suspected in precision mental health: there is no one feature that explains sufficient variance to make differential predictions by treatment on its own. Rather, we need to consider many features simultaneously, each of which offers only a small contribution but which together can guide treatment selection.

All the features considered in this model were expected to be relevant to predicting smoking cessation overall. The selection of baseline characteristics to be assessed in the original trial was guided by domain expertise and decades of research (Baker et al., 2016). For example, it makes sense that having most of your friends or family smoke would decrease treatment success. What may be less immediately intuitive, however, is why this characteristic further decreases treatment success specifically when treated with varenicline. This example demonstrates the value of the suite of machine learning tools: models can be built with high-dimensional data, allowing us to identify unexpected relationships, which we can then elucidate and understand using interpretable machine learning techniques.

These non-intuitive prescriptive factors also highlight the potential downsides of a highly transparent, interpretable model. When patients and clinicians can easily see non-intuitive features, they may be dissuaded from using or trusting this treatment selection model. It is important to remember that because GLMNet aims to reduce dimensionality by removing highly correlated features, a non-intuitive feature may be retained that is highly correlated with a more intuitive feature. For example, we cannot explain why an interaction feature with the ASI-3 item "When my throat is tight, I worry I will choke to death" was retained. However, it is possible that this feature was highly correlated with many other features related to anxiety about physical conditions. Because it seems to indicate a rather severe level of this specific type of anxiety, perhaps it offered more predictive power than other correlated features. Regardless of the possible underlying reason, the reality is that a patient will see this question on their treatment selection assessment, and it may make them doubt the trustworthiness of the model.

Prognostic factors for treatment development

In addition to prescriptive factors, we identified *prognostic* factors that predict treatment success overall (i.e., features with "main effects"). Although these features are not useful for selecting among the treatments in our model, they are valuable in several other ways.

These features contribute to the literature on prognostic factors that predict treatment success in the area of cigarette smoking. In particular, they support the conclusion from a recent review that predictors of treatment success span many categories (Bickel et al., 2023). We found similar breadth in important features in this model: economic (e.g., income), environmental (e.g., living with another smoker), sociodemographic (e.g., marital status), psychological (e.g., depression diagnosis), physical health (e.g., pain interfering with daily activities), and smoking use/history (e.g., longest previous quit attempt) characteristics all contributed to predicting smoking cessation.

Additionally, these prognostic factors may yet help to advance precision mental health goals. These factors may represent mechanisms underlying smoking cessation success and thus offer targeted areas for future treatment development. They also may be used to tailor existing treatments to increase success across individuals. For example, greater life satisfaction and higher importance of quitting predicted greater treatment success. Perhaps motivational interviewing/motivational enhancement therapy techniques that might address these factors could be used in pre-cessation counseling. These improvements could be made more scalable still by offering psychoeducation and support tool links in a website where people complete the remote assessment for treatment selection.

It is also possible that features that emerged as prognostic factors in our model may serve as prescriptive factors related to other treatments not included in this study. For example, buproprion is another first-line smoking cessation medication (Cahill et al., 2013). It may be that some features in this model do not differentiate among C-NRT, nicotine patch, or varenicline but would differentiate between one of these treatments and buproprion.

The role of demographic features

Several demographic features emerged as important prognostic and prescriptive factors: race, gender, income, and marital status. Ethnicity did not emerge as an important feature, and race-based features were specifically related to identifying as a Black or White individual. However, the limited representation of Hispanic, Latino/a, Asian, Multiracial, and Native American/Alaska Native individuals in this sample warrants caution in drawing conclusions about the predictive utility of ethnicity- and race-based features.

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 is discriminatory (e.g., (Farayola et al., 2023)). 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.

Future directions

Algorithmic fairness

Our sample was representative across several demographic characteristics; however, as noted above, we had poor representation of several racial minority groups and of Hispanic and Latino/a individuals. Our model can only be as good as the data with which it was developed (Aldridge, 2019). We must make a concerted effort to build treatment selection models using data where there was good representation across as many marginalized characteristics as possible. Otherwise, we run the risk of exacerbating rather than mitigating mental healthcare disparities.

There are also tools to examine the fairness of a prediction algorithm across sub-populations. For example, we could assess whether our model performs as well for White and non-White individuals. If there are differences, we can also dig deeper into understanding *why* our models perform differently by examining the differences in feature utility across these same groups. We could similarly examine whether our treatment selection benefit differs by any demographic characteristics in our sample. We plan to pursue both these analyses to identify biases at play in our prediction and treatment selection models.

Prospective clinical trial

We made a concerted effort in this project to evaluate how our treatment selection model would perform for new patients. Specifically, we used leave-one-out cross-validation to calculate predictions for each held-out individual that were then used to select that person's best treatment.

Regardless, the ultimate test of this model's clinical benefit will be in a prospective trial. This trial will offer two tests. First, it will assess whether using this model is feasible and acceptable to patients and clinicians in clinical practice. There may be opportunities to incorporate input directly from these stakeholders. Second, we can evaluate the benefit of our treatment selection model in an entirely new sample. Individuals who receive a model-assigned best treatment could be compared to any one of several possible comparison groups. Individuals in the comparison group could receive a random treatment assignment (mimicking clinical trials). Alternatively, they could receive clinician-assigned treatment to mirror traditional treatment selection (and best current clinical practice). Another option is that we could compare to a simpler model. For example, there is evidence that treatment success is lower when people are re-treated with the same treatment (Fiore et al., 2008; Gonzales et al., 2014; Heckman et al., 2017; Tønnesen et al., 1993) and that treatment adherence is higher when people choose their preferred treatment (Cropsey et al., 2017). Thus, treatments could be assigned based on patient preference and re-treatment status. Each comparison offers different advantages and disadvantages that should be considered thoughtfully when designing a prospective trial.

Conclusion

Overall, this study has potential for immediate benefit to individuals looking to quit smoking. Our treatment selection model can improve the probability of abstinence during early recovery by a statistically significant and clinically meaningful margin. Moreover, it can do so with a relatively low-burden assessment that uses widely accessible features. The ultimate test of this treatment selection model will be in a prospective trial that assesses the feasibility, acceptability, and effectiveness of this tool in clinical practice. We are optimistic about the promise our model holds to improve the public health burden of cigarette smoking.

Appendix 1: Supplemental Methods

AIM 1 analytic strategy

Model building

We built supplemental models to predict treatment success at 12 weeks and 26 weeks as measured via biologically confirmed, 7-day point-prevalence abstinence. The 12-week outcome represents the end-of-treatment and has been used as a primary outcome for this reason in extant precision mental health research (e.g., (Chen et al., 2020)). The 26-week (6 month) outcome is a typical outcome used in smoking cessation research to evaluate treatments because it serves as a feasible proxy for long-term abstinence (Fiore et al., 2008).

We followed our model fitting, selection, and evaluation procedures from the main manuscript to fit these additional models. Briefly, we considered model configurations that used the GLMNet statistical algorithm and varied by hyperparameter values and feature sets. We used nested cross-validation with 1 repeat of 10-fold cross-validation in the inner loop and 3 repeats of 10-fold cross-validation in the outer loop.

Metrics

Models were evaluated using area under the Receiver Operating Characteristic Curve (auROC) from held-out folds (test sets) in the outer loop. We used the same satisficing criterion of retaining a median of 50 or more treatment interactions across inner folds. We opted to keep this threshold consistent across models, though we confirmed that this value was still reasonable based on inner fold distributions and number of remaining model configurations for selection in each outer fold.

Bayesian analysis of model performance

We followed our same procedure to evaluate model performance using Bayesian hierarchical linear models. We estimated posterior probability distributions and 95% Bayesian credible intervals (CIs) following recommendations from the tidymodels team (Kuhn, 2022). 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 examined whether our models' performance differed as a function of prediction outcome. We regressed the auROCs (logit transformed) from the 30 test sets as a function of prediction outcome (4 week, 12 week, 26 week) with two random intercepts for repeat and fold within repeat. We report the 95% (equal-tailed) Bayesian CIs from the posterior probability distributions for the difference in performance between our models. If the 95% Bayesian CIs around the model contrast do not include 0, we can conclude that the models' performance differs by prediction outcome.

AIM 2 analytic strategy

We followed our methods from the main manuscript to select final model configurations, identify model-predicted best treatment using leave-one-out cross-validation, categorize treatment matching, and evaluate clinical benefit of these treatment selection models.

Although we did not preregister completing these analyses with 12-week and 26-week models, we followed our preregistered analyses. Like in our primary analyses, we used log base 2 for our log transformation of week instead of our preregistered log base e due to convergence issues.

Appendix 2: Supplemental Results

AIM 1 results: Prediction models

Model performance

We selected the best model configurations using auROCs from the *validation sets* (among models that met our satisficing metric of a median of at least 50 retained treatment interaction features in inner folds). We evaluated these best model configurations using *test set* performance. Test set performance for the 4-week model appears in the main manuscript. The median auROC across the 30 test sets for the 12-week model was 0.665 (IQR = 0.628 - 0.697, range = 0.535 - 0.788). The median auROC across the 30 test sets for the 26-week model was 0.629 (IQR = 0.573 - 0.658, range

= 0.474 - 0.743). The single, concatenated ROC curve and the 30 individual ROC curves (one per held-out fold) for the 4-, 12-, and 26-week models appear in Figure 5.

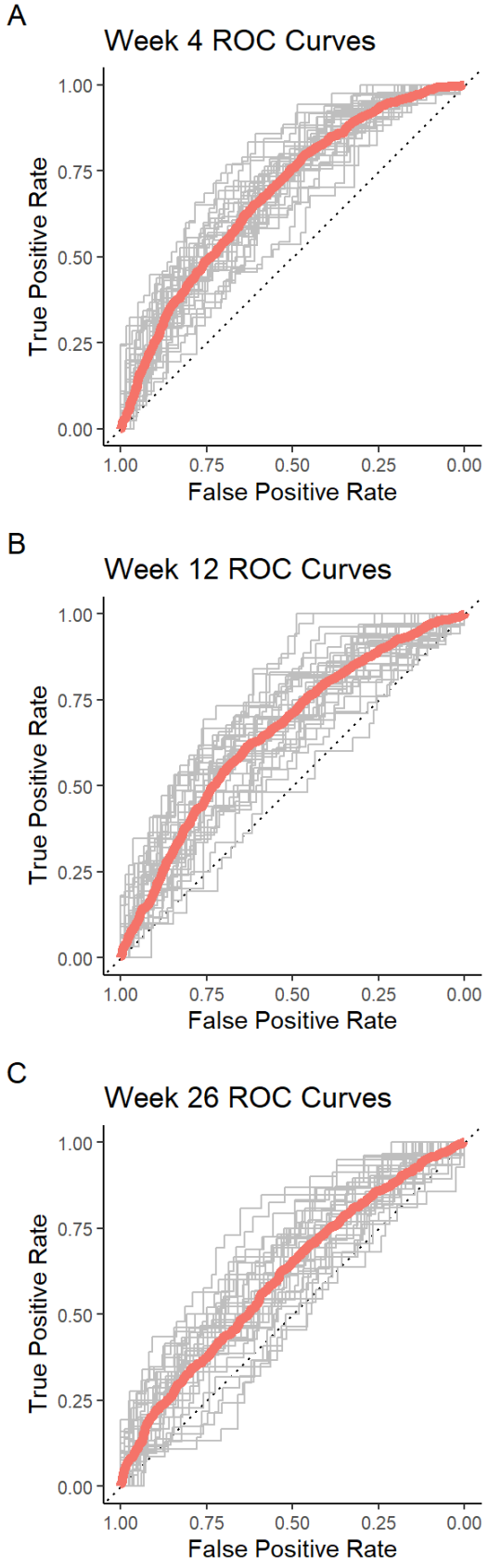
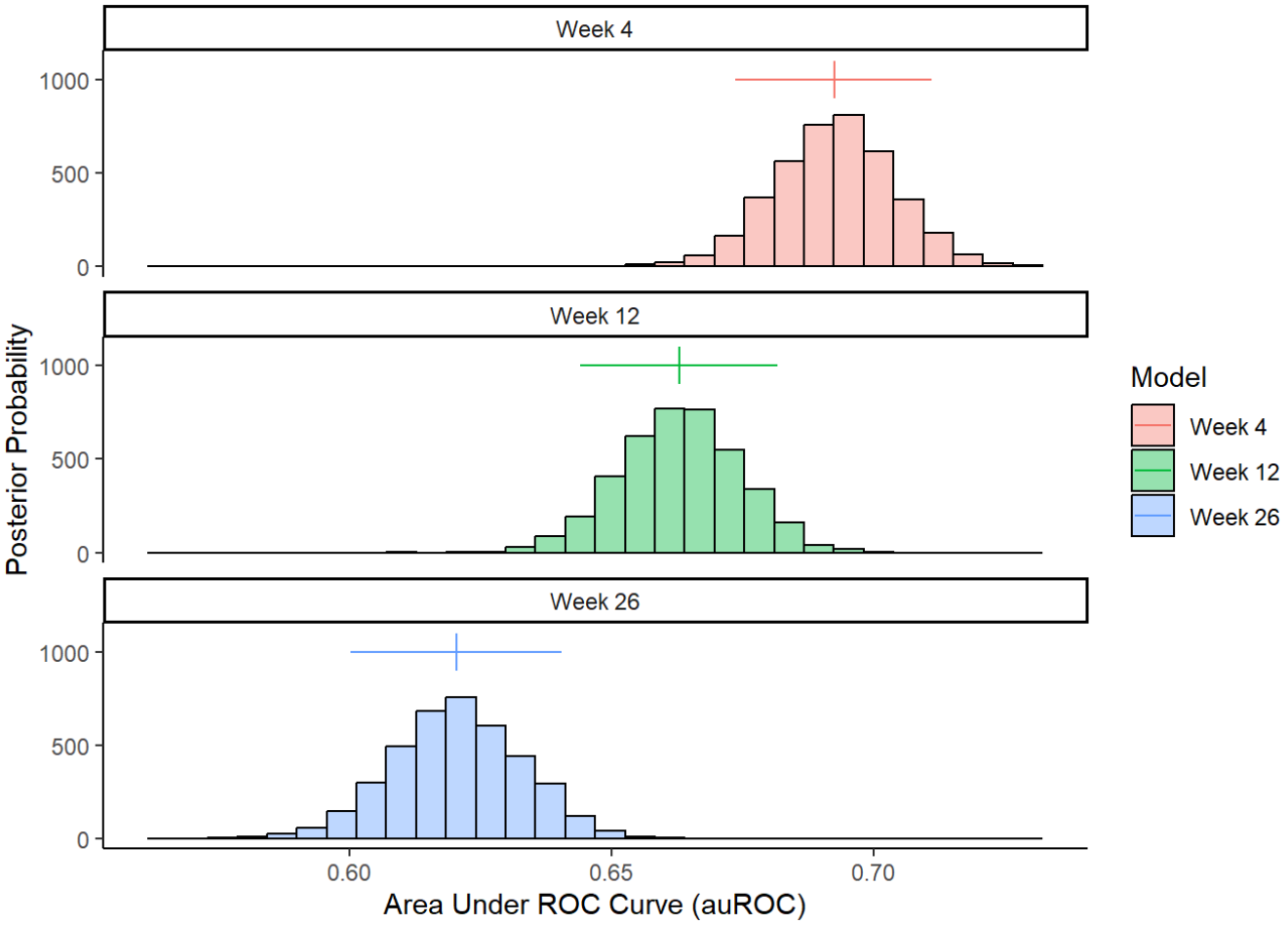


Figure 5: (Supplemental) ROC Curves. Dotted, diagonal line represents chance performance (0.5). Grey lines display individual ROC curves from each of 30 held-out folds. Thick, red line displays ROC curve concatenated across all 30 held-out folds. A) 4-week model. B) 12-week model. C) 26-week model.

We used the 30 test set auROCs to estimate the posterior probability distribution for the auROC of these models. The median auROC from the posterior distribution was 0.663 [Bayesian CI: 0.640 - 0.685] for the 12-week model and 0.620 [Bayesian CI: 0.596 - 0.644] for the 26-week model. These results suggest both models have predictive signal as the CIs did not contain 0.5 (chance performance). Figure 6 displays the posterior probability distributions for these models' auROCs alongside the posterior probability distribution for the 4-week model's auROCs.

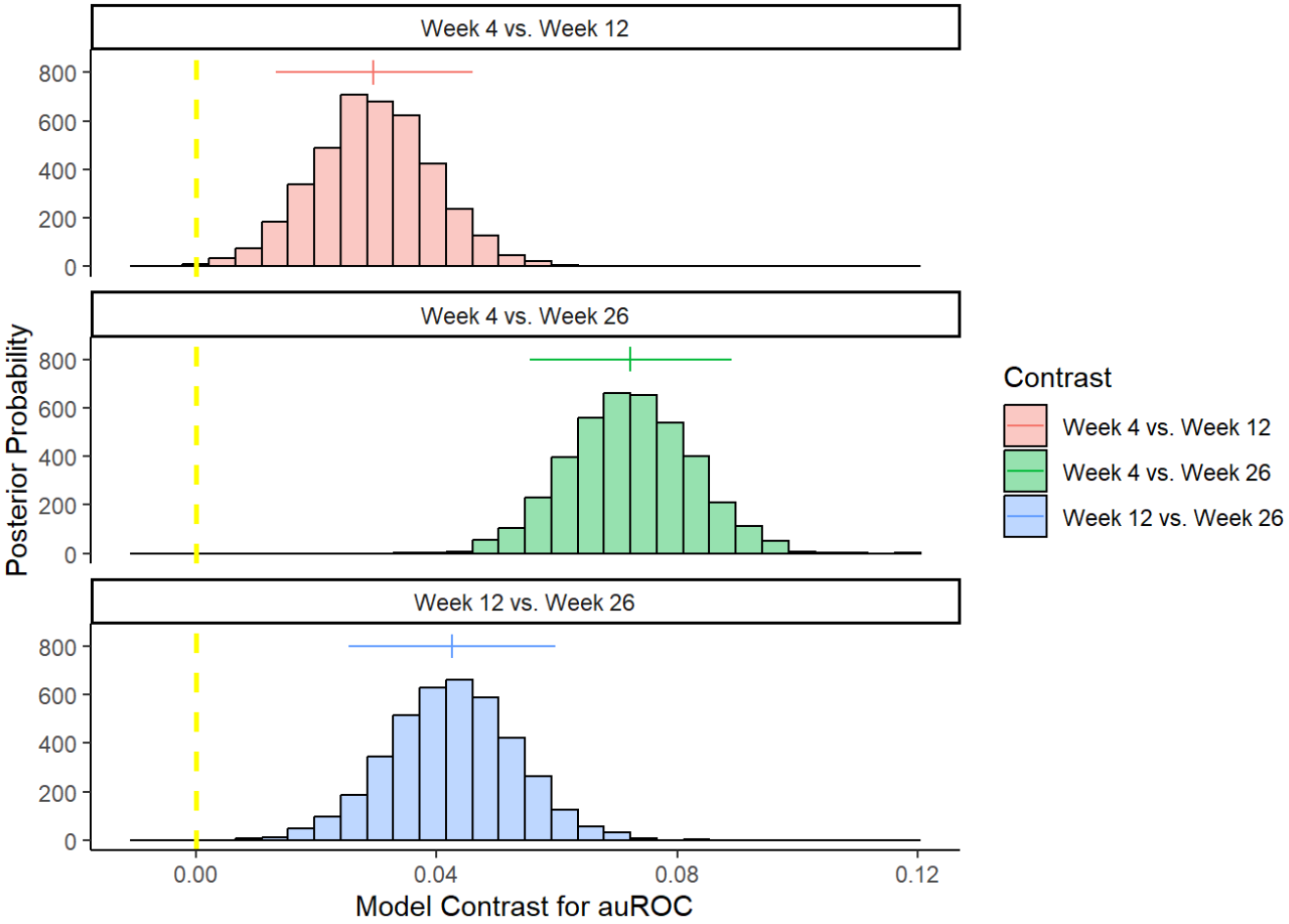
Figure 6: (Supplemental) Posterior probability distributions for the median auROC in test sets. Histogram represents posterior probability distribution. Horizontal line displays 95% Bayesian credible interval.



Model comparisons

We used the posterior probability distributions for the auROCs to compare the 4-, 12-, and 26-week models. The median increase in auROC for the 4- vs. 12-week model was 0.029 (95% CI = 0.010 - 0.049), yielding a probability of 99.8% that the 4-week model had superior performance. The median increase in auROC for the 4- vs. 26-week model was 0.072 (95% CI = 0.052 - 0.092), yielding a probability of 100% that the 4-week model had superior performance. The median increase in auROC for the 12- vs. 26-week model was 0.043 (95% CI = 0.022 - 0.064), yielding a probability of 100% that the 12-week model had superior performance. Figure 7 presents histograms of the posterior probability distributions for these model contrasts.

Figure 7: (Supplemental) Posterior probability distributions for the model contrasts. Histogram represents posterior probability distribution. Horizontal line displays 95% Bayesian credible interval.



Model calibration

In Figure 8, we display the calibration for the 4-week model (reproduced from main text; Panel A), the 12-week model (Panel B) and 26-week model (Panel C). Predicted lapse probabilities are binned (bin width = 10%) and plotted against the observed probability of abstinence for observations in that bin. If probabilities were perfectly calibrated, all bin means would fall on the dotted line (e.g., bin from 0 - 10 with an observed mean probability of 0.05, bin from 10 - 20 with an observed mean probability of 0.15).

Each figure plots the probabilities from our held-out predictions (made with leave-one-out cross-validation) using the final selected model configuration for each individual for their original trial-assigned treatment against the observed trial abstinence rates. Probabilities were relatively well calibrated and ordinal in their relationship with the true probability of abstinence for both the 12- and 26-week models. Given this, these probabilities can provide precise predictions of treatment success that can be used for treatment selection.

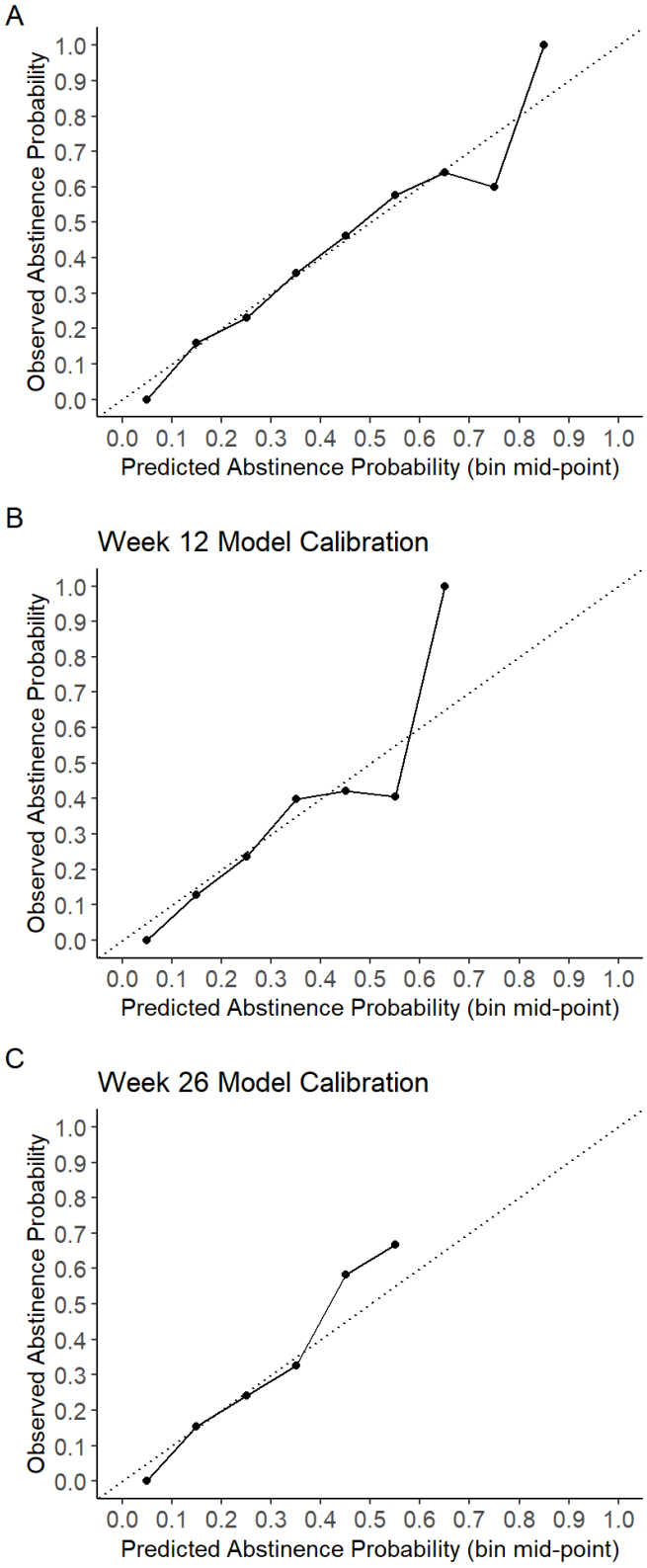


Figure 8: (Supplemental) Binned calibration plots. A) 4-week model. B) 12-week model. C) 26-week model.

Model interpretation

The names of all retained features and their parameter estimates from the final 4-week model appear in Table 5. Retained features for the 4-week model are discussed in detail in the main manuscript.

Table 5: (Supplemental) Retained features in 4-week final model

term	estimate
(Intercept)	−0.7492
co	−0.1824
race_ehr_white	0.1280
treatment_combo_nrt_x_live_with_smoker_live_alone_or_only_with_partner	0.1076
wisdm37_2_item_order	0.1062
motive_quit_order	0.1041
asi3_15_item_order	0.0896
wisdm37_30_item_order	−0.0893
longest_quit_order	0.0836
wsws_want_cig_24h_item_order	−0.0769
treatment_varenicline_x_live_with_smoker_live_alone_or_only_with_partner	−0.0722
race_ehr_black_african_american	−0.0675
shp_2_item_order	0.0673
dts_13_item_order	0.0669
treatment_combo_nrt_x_race_ehr_black_african_american	−0.0663
time_around_smokers_weekend_order	−0.0645
life_satisfaction_order	0.0615
treatment_combo_nrt_x_wisdm37_22_item_order	−0.0607
income_order	0.0557
mfi_15_item_order	0.0518
mfi_9_item_order	−0.0478
treatment_combo_nrt_x_smoke_menthol_no	0.0468
tot_quit_attempt	0.0466
wisdm37_5_item_order	−0.0456
important_to_quit_order	0.0431
treatment_patch_x_used_cigars_never_tried	−0.0428
treatment_patch_x_close_smoke_friend_no	−0.0419
treatment_combo_nrt_x_dts_5_item_order	0.0415
wisdm37_27_item_order	−0.0414
smoke_menthol_yes	−0.0403
wisdm37_13_item_order	−0.0389
shp_9_item_order	0.0384
treatment_varenicline_x_asi3_15_item_order	0.0381
[illegible]	[illegible]

Our final 12-week model fit with the full dataset retained retained 111 features (Table 6). Of these, 47 were treatment interaction features. Although we required that model configurations have a median of 50 or more treatment interaction terms retained across configurations, there was variability among folds; thus, it is unsurprising that the final model may have slightly fewer retained interaction features.

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 29 unique items (e.g., multiple dummy variables are from a single item, the same feature interacts with more than one treatment).

Table 6: (Supplemental) Retained features in 12-week final model

term	estimate
(Intercept)	−0.9527
co	−0.1268
race_ehr_white	0.1130
wisdm37_en_goad_scale	−0.0791
treatment_varenicline_x_confident_quit_order_09	0.0730
age_ehr	0.0696
wsws_crave_scale	−0.0677
alc_qty_order_09	0.0571
race_ehr_black_african_american	−0.0569
alc_freq_order_02	0.0524
treatment_varenicline_x_income_order_1	−0.0522
longest_quit_order_2	−0.0496
life_satisfaction_order_2	−0.0473
confident_quit_order_09	0.0470
time_around_smokers_order_1	0.0465
income_order_1	−0.0464
motive_quit_order_10	0.0461
treatment_patch_x_time_around_smokers_order_3	−0.0431
last_rct_quit_order_7	−0.0421
treatment_varenicline_x_marital_status_ehr_divorced	−0.0419
treatment_combo_nrt_x_alc_freq_order_10	0.0409
wisdm37_tolerance_scale	−0.0408
smoke_menthol_no	0.0374
sip2r_total_scale	−0.0360
shp_total_scale	0.0358
treatment_varenicline_x_berlin_7_order_2	0.0332
employment_unemployed	−0.0327
treatment_combo_nrt_x_smoke_menthol_yes	−0.0327
time_around_smokers_order_3	−0.0323
treatment_combo_nrt_x_berlin_7_order_5	0.0315
alc_freq_order_05	−0.0314
treatment_patch_x_tried_chantix_yes	−0.0309
smoke_menthol_yes	−0.0303
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]

Our final 26-week model fit with the full dataset retained retained 107 features (Table 7). Of these, 45 were treatment interaction features. Like above, variability among folds led to the final model having slightly fewer than 50 retained interaction features.

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 36 unique items (e.g., multiple dummy variables are from a single item, the same feature interacts with more than one treatment).

Table 7: (Supplemental) Retained features in 26-week final model

term	estimate
(Intercept)	−1.1683
co	−0.1460
wisdm37_en_goad_scale	−0.0662
motive_quit_order	0.0648
shp_total_scale	0.0647
race_ehr_white	0.0631
longest_quit_order	0.0576
tot_quit_attempt	0.0542
treatment_combo_nrt_x_spouse_smoke_dont_live_with_spouse_or_partner	0.0529
treatment_varenicline_x_marital_status_ehr_divorced	−0.0528
wsws_neg_aff_scale	−0.0522
wisdm37_tolerance_scale	−0.0489
smoke_menthol_yes	−0.0478
age_1st_cig	−0.0471
dsm5_score_scale	−0.0458
treatment_combo_nrt_x_live_with_smoker_live_alone_or_only_with_partner	0.0424
wsws_crave_scale	−0.0419
race_ehr_black_african_american	−0.0375
age_ehr	0.0342
hdsm_tired	−0.0332
wisdm37_cue_scale	−0.0322
wisdm37_auto_scale	−0.0317
panas_neg_scale	−0.0271
marital_status_ehr_divorced	−0.0271
wsws_hunger_scale	0.0268
treatment_varenicline_x_tried_nic_gum_yes	0.0263
time_around_smokers_order	−0.0251
smoke_menthol_no	0.0251
tried_nic_gum_yes	0.0243
tried_nic_gum_no	−0.0243
treatment_combo_nrt_x_asi3_cognitive_scale	0.0238
treatment_varenicline_x_income_order	0.0228
treatment_varenicline_x_tried_nic_gum_no	−0.0223
[illegible]	[illegible]

AIM 2 results: Clinical benefit

There was no significant fixed effect of treatment matching for the 12-week ($p = 0.279$) or the 26-week ($p = 0.967$) model. The treatment matching X time interaction was also not significant for the 12-week ($p = 0.660$) or the 26-week ($p = 0.402$) model. There was a significant fixed effect of time in both models (12-week model: OR $= 0.210$, $z = -9.973$, $p < 0.001$; 26-week model: OR $= 0.208$, $z = -10.050$, $p < 0.001$) such that the probability of treatment success declined over time. Mean treatment success by treatment matching over time from the 4-week model is reproduced for visual comparison purposes in Figure 9. Figure 10 and Figure 11 follow the same figure format for the 12- and 26-week models, respectively.

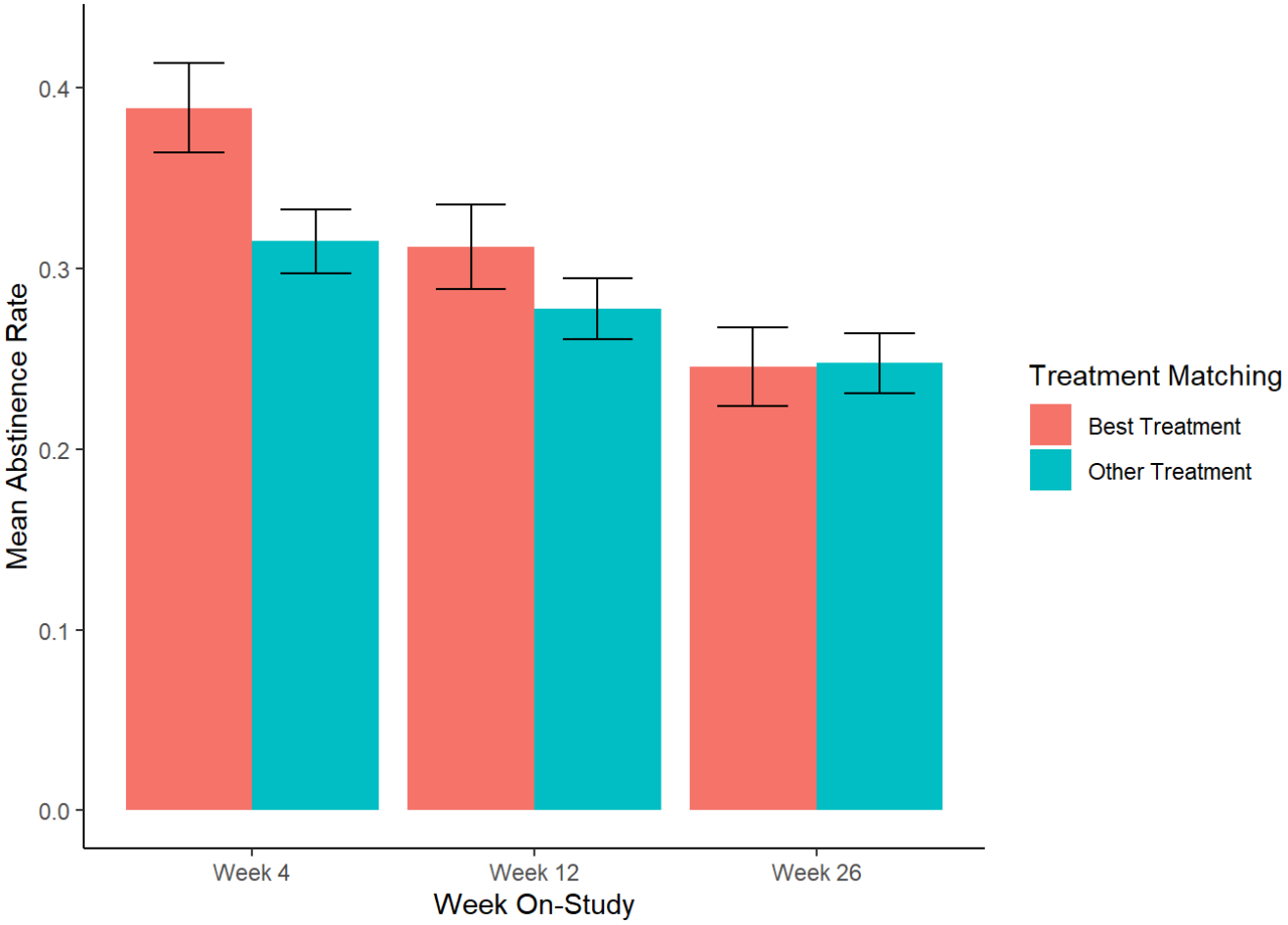


Figure 9: (Supplemental; reproduced from main text) Benefit of treatment matching from 4-week prediction model. Bars represent mean observed abstinence (from original trial) for individuals who did and did not receive their model-predicted best treatment, over time. Error bars indicate standard errors.

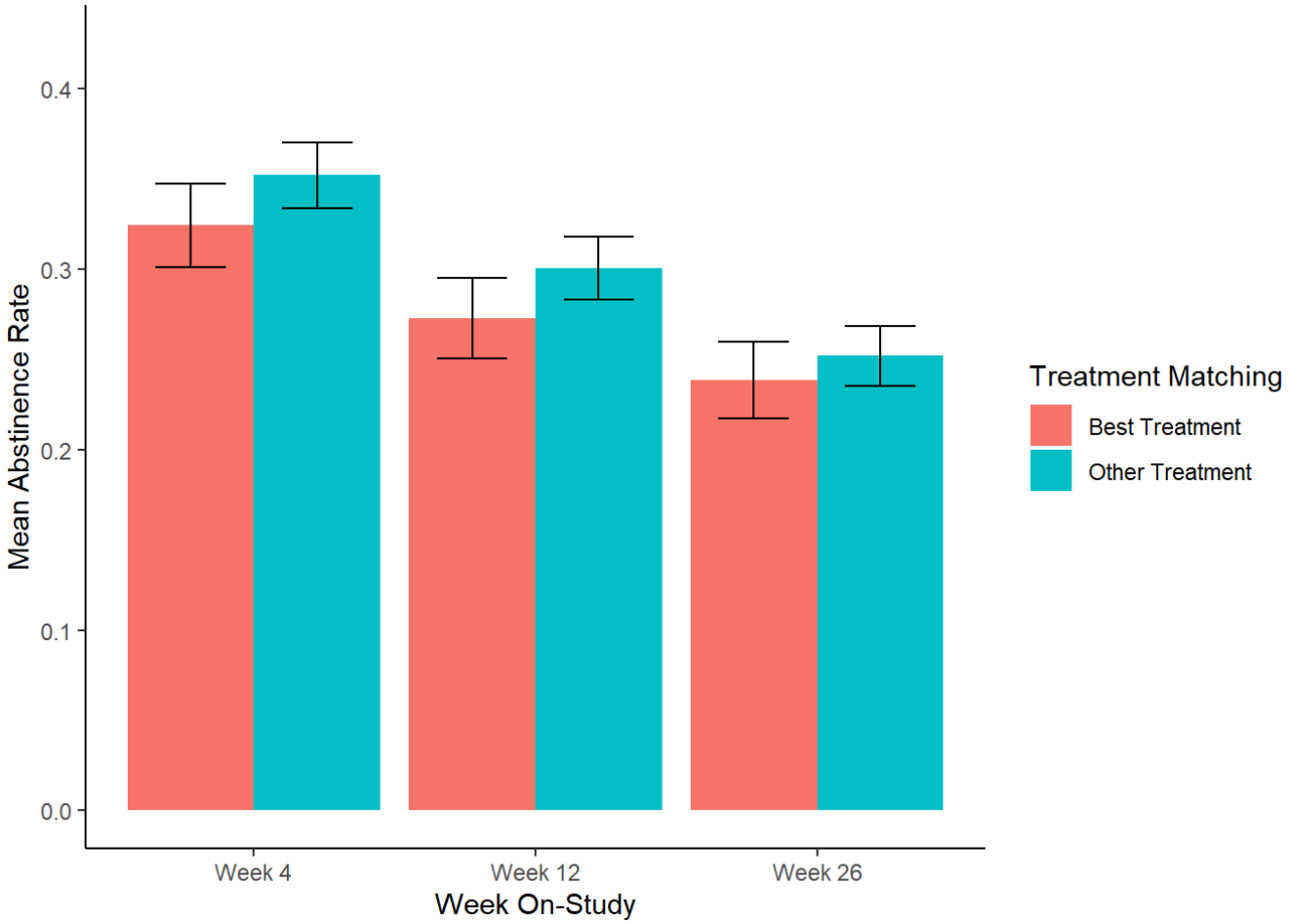


Figure 10: (Supplemental) Benefit of treatment matching from 12-week prediction model. Bars represent mean observed abstinence (from original trial) for individuals who did and did not receive their model-predicted best treatment, over time. Error bars indicate standard errors.

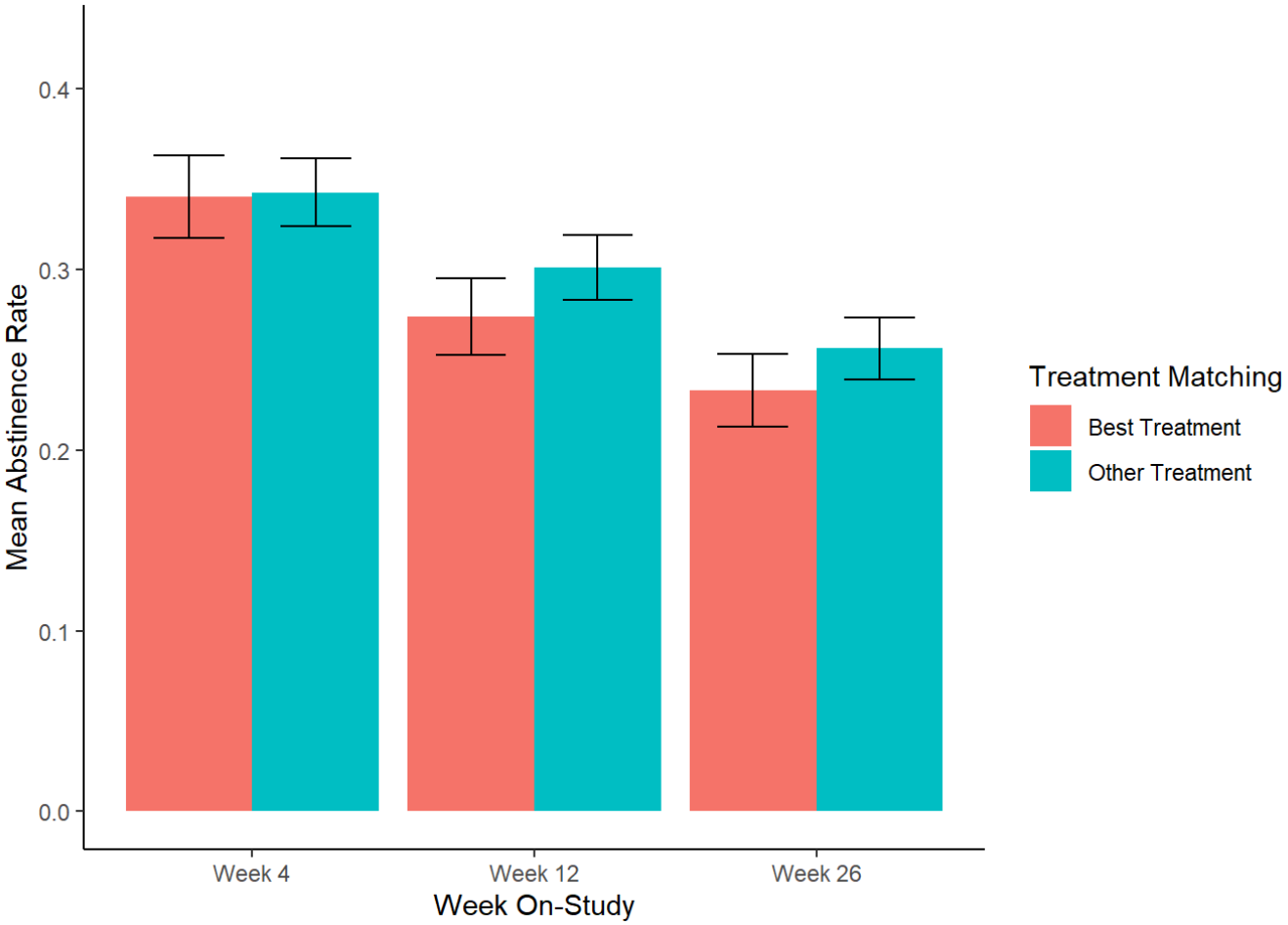


Figure 11: (Supplemental) Benefit of treatment matching from 12-week prediction model. Bars represent mean observed abstinence (from original trial) for individuals who did and did not receive their model-predicted best treatment, over time. Error bars indicate standard errors.

Bibliography

(2016).

Adjei, K., & Ali, A. A. (2022). CO110 Comparative Effectiveness of Sertraline, Fluoxetine Vs Escitalopram Among Adults with Depression in the United States. *Value in Health*, *25*(7), S324-S325. https://doi.org/10.1016/j.jval.2022.04.206

Aldridge, R. W. (2019). Research and Training Recommendations for Public Health Data Science. *The Lancet. Public Health*, *4*(8), e373. https://doi.org/10.1016/S2468-2667(19)30112-4

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition). American Psychiatric Association. https://doi.org/10.1176/appi.books.9780890425596

Bae, S., Chung, T., Ferreira, D., Dey, A. K., & Suffoletto, B. (2018). Mobile Phone Sensors and Supervised Machine Learning to Identify Alcohol Use Events in Young Adults: Implications for Just-in-Time Adaptive Interventions. *Addictive Behaviors*, *83*, 42–47. https://doi.org/10.1016/j.addbeh.2017.11.039

Baggett, T. P., Tobey Matthew L., & Rigotti Nancy A. (2013). Tobacco Use among Homeless People — Addressing the Neglected Addiction. *New England Journal of Medicine*, *369*(3), 201–204. https://doi.org/10.1056/NEJMp1301935

Baker, T. B., & McCarthy, D. E. (2021). Smoking Treatment: A Report Card on Progress and Challenges. *Annual Review of Clinical Psychology*, *17*(Volume17, 2021), 1–30. https://doi.org/10.1146/annurev-clinpsy-081219-090343

Baker, T. B., Piper, M. E., Stein, J. H., Smith, S. S., Bolt, D. M., Fraser, D. L., & Fiore, M. C. (2016). Effects of Nicotine Patch vs Varenicline vs Combination Nicotine Replacement Therapy on Smoking Cessation at 26 Weeks: A Randomized Clinical Trial. *JAMA*, *315*(4), 371–379. https://doi.org/10.1001/jama.2015.19284

Barksdale, C. L., Pérez-Stable, E., & Gordon, J. (2022). Innovative Directions to Advance Mental Health Disparities Research. *American Journal of Psychiatry*, *179*(6), 397–401. https://doi.org/10.1176/appi.ajp.21100972

Bickel, W. K., Tomlinson, D. C., Craft, W. H., Ma, M., Dwyer, C. L., Yeh, Y.-H., Tegge, A. N., Freitas-Lemos, R., & Athamneh, L. N. (2023). Predictors of Smoking Cessation Outcomes Identified by Machine Learning: A Systematic Review. *Addiction Neuroscience*, *6*, 100068. https://doi.org/10.1016/j.addicn.2023.100068

Bickman, L. (2020). Improving Mental Health Services: A 50-Year Journey from Randomized Experiments to Artificial Intelligence and Precision Mental Health. *Administration and Policy in Mental Health*, 1–49. https://doi.org/10.1007/s10488-020-01065-8

Bickman, L., Lyon, A. R., & Wolpert, M. (2016). Achieving Precision Mental Health through Effective Assessment, Monitoring, and Feedback Processes. *Administration and Policy in Mental Health and Mental Health Services Research*, *43*(3), 271–276. https://doi.org/10.1007/s10488-016-0718-5

Biecek, P. (2018). DALEX: Explainers for Complex Predictive Models in R. *Journal of Machine Learning Research*, *19*(84), 1–5.

Bogdan, R., Baranger, D. A., & Agrawal, A. (2018). Polygenic Risk Scores in Clinical Psychology: Bridging Genomic Risk to Individual Differences. *Annual Review of Clinical Psychology*, *14*(1), 119–157. https://doi.org/10.1146/annurev-clinpsy-050817-084847

Brandon, T. H., Vidrine, J. I., & Litvin, E. B. (2007). Relapse and Relapse Prevention. *Annual Review of Clinical Psychology*, *3*(1), 257–284. https://doi.org/10.1146/annurev.clinpsy.3.022806.091455

Cahill, K., Lindson-Hawley, N., Thomas, K. H., Fanshawe, T. R., & Lancaster, T. (2016). Nicotine Receptor Partial Agonists for Smoking Cessation. *The Cochrane Database of Systematic Reviews*, *2016*(5). https://doi.org/10.1002/14651858.CD006103.pub7

Cahill, K., Stevens, S., Perera, R., & Lancaster, T. (2013). Pharmacological Interventions for Smoking Cessation: An Overview and Network Meta-Analysis. *The Cochrane Database of Systematic Reviews*, *5*, CD9329. https://doi.org/10.1002/14651858.CD009329.pub2

Centers for Disease Control and Prevention (CDC). *Annual Average for United States 2011–2015 Alcohol-Attributable Deaths Due to Excessive Alcohol Use, All Ages.*

Chen, L.-S., Baker, T. B., Miller, J. P., Bray, M., Smock, N., Chen, J., Stoneking, F., Culverhouse, R. C., Saccone, N. L., Amos, C. I., Carney, R. M., Jorenby, D. E., & Bierut, L. J. (2020). Genetic Variant in CHRNA5 and Response to Varenicline and Combination Nicotine Replacement in a Randomized Placebo-Controlled Trial. *Clinical Pharmacology & Therapeutics*, *108*(6), 1315–1325. https://doi.org/10.1002/cpt.1971

Chen, L.-S., Horton, A., & Bierut, L. (2018). Pathways to Precision Medicine in Smoking Cessation Treatments. *Neuroscience Letters*, *669*, 83–92. https://doi.org/10.1016/j.neulet.2016.05.033

Chenoweth, M. J., Schnoll, R. A., Novalen, M., Hawk, L. W., George, T. P., Cinciripini, P. M., Lerman, C., & Tyndale, R. F. (2016). The Nicotine Metabolite Ratio Is Associated With Early Smoking Abstinence Even After Controlling for Factors That Influence the Nicotine Metabolite Ratio. *Nicotine & Tobacco Research*, *18*(4), 491–495. https://doi.org/10.1093/ntr/ntv125

Chih, M.-Y., Patton, T., McTavish, F. M., Isham, A. J., Judkins-Fisher, C. L., Atwood, A. K., & Gustafson, D. H. (2014). Predictive Modeling of Addiction Lapses in a Mobile Health Application. *Journal of Substance Abuse Treatment*, *46*(1), 29–35. https://doi.org/10.1016/j.jsat.2013.08.004

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A Nondegenerate Penalized Likelihood Estimator for Variance Parameters in Multilevel Models. *Psychometrika*, *78*(4), 685–709. https://doi.org/10.1007/s11336-013-9328-2

Cohen, Z. D., & DeRubeis, R. J. (2018). Treatment Selection in Depression. *Annual Review of Clinical Psychology*, *14*(1), 209–236. https://doi.org/10.1146/annurev-clinpsy-050817-084746

Collins, L. M. (2018). *Optimization of Behavioral, Biobehavioral, and Biomedical Interventions: The Multiphase Optimization Strategy (MOST)*. Springer International Publishing. https://doi.org/10.1007/978-3-319-72206-1

Cornelius, M. E. (2020). Tobacco Product Use Among Adults — United States, 2019. *MMWR. Morbidity and Mortality Weekly Report*, *69*. https://doi.org/10.15585/mmwr.mm6946a4

Coughlin, L. N., Tegge, A. N., Sheffer, C. E., & Bickel, W. K. (2020). A Machine-Learning Approach to Predicting Smoking Cessation Treatment Outcomes. *Nicotine & Tobacco Research*, *22*(3), 415–422. https://doi.org/10.1093/ntr/nty259

Crawford, J. R., & Henry, J. D. (2004). The Positive and Negative Affect Schedule (PANAS): Construct Validity, Measurement Properties and Normative Data in a Large Non-Clinical Sample. *British Journal of Clinical Psychology*, *43*(3), 245–265. https://doi.org/10.1348/0144665031752934

Cropsey, K. L., Hendricks, P. S., Schiavon, S., Sellers, A., Froelich, M., Shelton, R. C., & Carpenter, M. J. (2017). A Pilot Trial of In Vivo NRT Sampling to Increase Medication Adherence in Community Corrections Smokers. *Addictive Behaviors*, *67*, 92–99. https://doi.org/10.1016/j.addbeh.2016.12.011

Cropsey, K., Eldridge, G. D., & Ladner, T. (2004). Smoking among Female Prisoners: An Ignored Public Health Epidemic. *Addictive Behaviors*, *29*(2), 425–431. https://doi.org/10.1016/j.addbeh.2003.08.014

DeRubeis, R. J. (2019). The History, Current Status, and Possible Future of Precision Mental Health. *Behaviour Research and Therapy*, *123*, 103506. https://doi.org/10.1016/j.brat.2019.103506

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., & Lorenzo-Luaces, L. (2014). The Personalized Advantage Index: Translating Research on Prediction into Individualized Treatment Recommendations. A Demonstration. *Plos One*, *9*(1), e83875. https://doi.org/10.1371/journal.pone.0083875

Dickson, M., & Gagnon, J. P. (2009). The Cost of New Drug Discovery and Development. *Discovery Medicine*, *4*(22), 172–179.

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine Learning Approaches for Clinical Psychology and Psychiatry. *Annual Review of Clinical Psychology*, *14*(1), 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Epstein, D. H., Tyburski, M., Kowalczyk, W. J., Burgess-Hull, A. J., Phillips, K. A., Curtis, B. L., & Preston, K. L. (2020). Prediction of Stress and Drug Craving Ninety Minutes in the Future with Passively Collected GPS Data. *Npj Digital Medicine*, *3*(1), 26. https://doi.org/10/ghqvcw

Etter, J.-F., Vera Cruz, G., & Khazaal, Y. (2023). Predicting Smoking Cessation, Reduction and Relapse Six Months after Using the Stop-Tabac App for Smartphones: A Machine Learning Analysis. *BMC Public Health*, *23*(1), 1076. https://doi.org/10.1186/s12889-023-15859-6

Farayola, M. M., Tal, I., Connolly, R., Saber, T., & Bendechache, M. (2023). Ethics and Trustworthiness of AI for Predicting the Risk of Recidivism: A Systematic Literature Review. *Information*, *14*(8), 426. https://doi.org/10.3390/info14080426

Feczko, E., & Fair, D. A. (2020). Methods and Challenges for Assessing Heterogeneity. *Biological Psychiatry*, *88*(1), 9–17. https://doi.org/10.1016/j.biopsych.2020.02.015

Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., Ripke, S., Day, F. R., Consortium, R., Purcell, S., Stahl, E., Lindstrom, S., Perry, J. R. B., Okada, Y., Raychaudhuri, S., … Price, A. L. (2015). Partitioning Heritability by Functional Annotation Using Genome-Wide Association Summary Statistics. *Nature Genetics*, *47*(11), 1228–1235. https://doi.org/10.1038/ng.3404

Fiore, M. C., Jaen, C. R., Baker, T. B., Bailey, W. C., Benowitz, N. L., Curry, S. J., Dorfman, S. F., Froelicher, E. S., Goldstein, M. G., Healton, C. G., Henderson, P. N., Heyman, R. B., Koh, H. K., Kottke, T. E., Lando, H. A., Mecklenburg, R. E., Mermelstein, R. J., Mullen, P. D., Orleans, C. T., … Wewers, M. E. (2008). *A Clinical Practice Guideline for Treating Tobacco Use and Dependence: 2008 Update. A U.S. Public Health Service Report.* Rockville, MD: U.S. Department of Health and Human Services, U.S. Public Health Service.

Gabry, J., & Goodrich, B. (2023). *Prior Distributions for Rstanarm Models.*

Glatard, A., Dobrinas, M., Gholamrezaee, M., Lubomirov, R., Cornuz, J., Csajka, C., & Eap, C. B. (2017). Association of Nicotine Metabolism and Sex with Relapse Following Varenicline and Nicotine Replacement Therapy. *Experimental and Clinical Psychopharmacology*, *25*(5), 353–362. https://doi.org/10.1037/pha0000141

Gonzales, D., Hajek, P., Pliamm, L., Nackaerts, K., Tseng, L.-J., McRae, T. D., & Treadow, J. (2014). Retreatment with Varenicline for Smoking Cessation in Smokers Who Have Previ-

ously Taken Varenicline: A Randomized, Placebo-Controlled Trial. *Clinical Pharmacology and Therapeutics*, *96*(3), 390–396. https://doi.org/10.1038/clpt.2014.124

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2023). *Rstanarm: Bayesian Applied Regression Modeling via Stan.*

Harrison, A., Ramo, D., Hall, S. M., Estrada-Gonzalez, V., & Tolou-Shams, M. (2020). Cigarette Smoking, Mental Health, and Other Substance Use among Court-Involved Youth. *Substance Use & Misuse*, *55*(4), 572–581. https://doi.org/10.1080/10826084.2019.1691593

Heatherton, T. F., Kozlowski, L. T., Frecker, R. C., & Fagerström, K. O. (1991). The Fagerström Test for Nicotine Dependence: A Revision of the Fagerström Tolerance Questionnaire. *British Journal of Addiction*, *86*(9), 1119–1127.

Heckman, B. W., Cummings, K. M., Kasza, K. A., Borland, R., Burris, J. L., Fong, G. T., McNeill, A., & Carpenter, M. J. (2017). Effectiveness of Switching Smoking-Cessation Medications Following Relapse. *American Journal of Preventive Medicine*, *53*(2), e63-e70. https://doi.org/10.1016/j.amepre.2017.01.038

Hu, Y., Lu, Q., Powles, R., Yao, X., Yang, C., Fang, F., Xu, X., & Zhao, H. (2017). Leveraging Functional Annotations in Genetic Risk Prediction for Human Complex Diseases. *PLOS Computational Biology*, *13*(6), e1005589. https://doi.org/10.1371/journal.pcbi.1005589

Insel, T. R. (2014). The NIMH Research Domain Criteria (RDoC) Project: Precision Medicine for Psychiatry. *American Journal of Psychiatry*, *171*(4), 395–397. https://doi.org/10.1176/appi.ajp.2014.14020138

Issabakhsh, M., Sánchez-Romero, L. M., Le, T. T. T., Liber, A. C., Tan, J., Li, Y., Meza, R., Mendez, D., & Levy, D. T. (2023). Machine Learning Application for Predicting Smoking Cessation among US Adults: An Analysis of Waves 1-3 of the PATH Study. *Plos One*, *18*(6), e286883. https://doi.org/10.1371/journal.pone.0286883

Jacobson, N. C., Kowatsch, T., & Marsch, L. A. (Eds.). (2022). *Digital Therapeutics for Mental Health and Addiction: The State of the Science and Vision for the Future* (1st edition). Academic Press.

Jamal, A., Homa, D. M., O'Connor, E., Babb, S. D., Caraballo, R. S., Singh, T., Hu, S. S., & King, B. A. (2015). Current Cigarette Smoking among Adults - United States, 2005-2014. *MMWR. Morbidity and Mortality Weekly Report*, *64*(44), 1233–1240. https://doi.org/10.15585/mmwr.mm6444a2

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R* (7th ed.). Springer-Verlag. https://doi.org/10.1007/978-1-4614-7138-7

Jha Prabhat, Ramasundarahettige Chinthanie, Landsman Victoria, Rostron Brian, Thun Michael, Anderson Robert N., McAfee Tim, & Peto Richard. (2013). 21st-Century Hazards of Smoking and Benefits of Cessation in the United States. *New England Journal of Medicine*, *368*(4), 341–350. https://doi.org/10.1056/NEJMsa1211128

Jonathan, P., Krzanowski, W. J., & McCarthy, W. V. (2000). On the Use of Cross-Validation to Assess Performance in Multivariate Prediction. *Statistics and Computing*, *10*(3), 209–229. https://doi.org/10.1023/A:1008987426876

Jordan, C. J., & Xi, Z.-X. (2018). Discovery and Development of Varenicline for Smoking Cessation. *Expert Opinion on Drug Discovery*, *13*(7), 671–683. https://doi.org/10.1080/17460441.2018.1458090

Kaiser, J. (2015). *Obama Gives East Room Rollout to Precision Medicine Initiative*.

Karaca-Mandic, P., Norton, E. C., & Dowd, B. (2012). Interaction Terms in Nonlinear Models. *Health Services Research*, *47*(1pt1), 255–274. https://doi.org/10.1111/j.1475-6773.2011.01314.x

Kaufmann, A., Hitsman, B., Goelz, P. M., Veluz-Wilkins, A., Blazekovic, S., Powers, L., Leone, F. T., Gariti, P., Tyndale, R. F., & Schnoll, R. A. (2015). Rate of Nicotine Metabolism and Smoking Cessation Outcomes in a Community-based Sample of Treatment-Seeking Smokers. *Addictive Behaviors*, *51*, 93–99. https://doi.org/10.1016/j.addbeh.2015.07.019

Kaye, J. T., Johnson, A. L., Baker, T. B., Piper, M. E., & Cook, J. W. (2020). Searching for Personalized Medicine for Binge Drinking Smokers: Smoking Cessation Using Varenicline, Nicotine Patch, or Combination Nicotine Replacement Therapy. *Journal of Studies on Alcohol and Drugs*, *81*(4), 426–435. https://doi.org/10.15288/jsad.2020.81.426

Kelly, P. J., Baker, A. L., Deane, F. P., Kay-Lambkin, F. J., Bonevski, B., & Tregarthen, J. (2012). Prevalence of Smoking and Other Health Risk Factors in People Attending Residential Substance Abuse Treatment. *Drug and Alcohol Review*, *31*(5), 638–644. https://doi.org/10.1111/j.1465-3362.2012.00465.x

Kessler, R. C., & Luedtke, A. (2021). Pragmatic Precision Psychiatry—A New Direction for Optimizing Treatment Selection. *JAMA Psychiatry*, *78*(12), 1384–1390. https://doi.org/10.1001/jamapsychiatry.2021.2500

Kiluk, B. D., Dreifuss, J. A., Weiss, R. D., Morgenstern, J., & Carroll, K. M. (2013). The Short Inventory of Problems – Revised (SIP-R): Psychometric Properties within a Large, Diverse Sample of Substance Use Disorder Treatment Seekers. *Psychology of Addictive Behaviors : Journal of the Society of Psychologists in Addictive Behaviors*, *27*(1), 307–314. https://doi.org/10.1037/a0028445

Kranzler, H. R., Smith, R. V., Schnoll, R., Moustafa, A., & Greenstreet-Akman, E. (2017). Precision Medicine and Pharmacogenetics: What Does Oncology Have That Addiction Medicine Does Not?. *Addiction*, *112*(12), 2086–2094. https://doi.org/10.1111/add.13818

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a Brief Depression Severity Measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Krstajic, D., Buturovic, L. J., Leahy, D. E., & Thomas, S. (2014). Cross-Validation Pitfalls When Selecting and Assessing Regression and Classification Models. *Journal of Cheminformatics*, *6*. https://doi.org/10.1186/1758-2946-6-10

Kuhn, M. (2022). *Tidyposterior: Bayesian Analysis to Compare Models Using Resampling Statistics*.

Kuhn, M., & Johnson, K. (2018). *Applied Predictive Modeling* (1st ed. 2013, Corr. 2nd printing 2018 edition). Springer. https://doi.org/10.1007/978-1-4614-6849-3

Kuhn, M., & Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models* (1edition ed.). Chapman and Hall/CRC.

Lai, C.-C., Huang, W.-H., Chang, B. C.-C., & Hwang, L.-C. (2021). Development of Machine Learning Models for Prediction of Smoking Cessation Outcome. *International Journal of Environmental Research and Public Health*, *18*(5), 2584. https://doi.org/10.3390/ijerph18052584

Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive Burden of Survey Questions and Response Times: A Psycholinguistic Experiment. *Applied Cognitive Psychology*, *24*(7), 1003–1020. https://doi.org/10.1002/acp.1602

Lerman, C., Schnoll, R. A., Hawk, L. W., Cinciripini, P., George, T. P., Wileyto, E. P., Swan, G. E., Benowitz, N. L., Heitjan, D. F., Tyndale, R. F., & PGRN-PNAT Research Group. (2015). Use of the Nicotine Metabolite Ratio as a Genetically Informed Biomarker of Response to Nicotine Patch or Varenicline for Smoking Cessation: A Randomised, Double-Blind Placebo-Controlled Trial. *The Lancet. Respiratory Medicine*, *3*(2), 131–138. https://doi.org/10.1016/S2213-2600(14)70294-2

Lewis, C., Roberts, N. P., Andrew, M., Starling, E., & Bisson, J. I. (2020). Psychological Therapies for Post-Traumatic Stress Disorder in Adults: Systematic Review and Meta-Analysis. *European Journal of Psychotraumatology*, *11*(1), 1729633. https://doi.org/10.1080/20008198.2020.1729633

Lieberman, J. A. (2004). Dopamine Partial Agonists: A New Class of Antipsychotic. *CNS Drugs*, *18*(4), 251–267. https://doi.org/10.2165/00023210-200418040-00005

Lindson, N., Chepkin, S. C., Ye, W., Fanshawe, T. R., Bullen, C., & Hartmann-Boyce, J. (2019). Different Doses, Durations and Modes of Delivery of Nicotine Replacement Therapy for Smoking Cessation. *Cochrane Database of Systematic Reviews*, *4*. https://doi.org/10.1002/14651858.CD013308

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777.

MacEachern, S. J., & Forkert, N. D. (2021). Machine Learning for Precision Medicine. *Genome*, *64*(4), 416–425. https://doi.org/10.1139/gen-2020-0131

Mandrekar, J. N. (2010). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer*, *5*(9), 1315–1316. https://doi.org/10.1097/JTO.0b013e3181ec173d

Marszalek, J. M., Barber, C., Kohlhart, J., & Cooper, B. H. (2011). Sample Size in Psychological Research over the Past 30 Years. *Perceptual and Motor Skills*, *112*(2), 331–348. https://doi.org/10.2466/03.11.PMS.112.2.331-348

Massago, M., Massago, M., Iora, P. H., Tavares Gurgel, S. J., Conegero, C. I., Carolino, I. D. R., Mushi, M. M., Chaves Forato, G. A., de Souza, J. V. P., Hernandes Rocha, T. A., Bonfim, S., Staton, C. A., Nihei, O. K., Vissoci, J. R. N., & de Andrade, L. (2024). Applicability of Machine Learning Algorithm to Predict the Therapeutic Intervention Success in Brazilian Smokers. *Plos One*, *19*(3), e295970. https://doi.org/10.1371/journal.pone.0295970

Mooney, S. J., & Pejaver, V. (2018). Big Data in Public Health: Terminology, Machine Learning, and Privacy. *Annual Review of Public Health*, *39*, 95–112. https://doi.org/10.1146/annurev-publhealth-040617-014208

Morales, D. A., Barksdale, C. L., & Beckel-Mitchener, A. C. (2020). A Call to Action to Address Rural Mental Health Disparities. *Journal of Clinical and Translational Science*, *4*(5), 463–467. https://doi.org/10.1017/cts.2020.42

Moriarty, D. (1996). CDC Studies Community Quality of Life. *NACCHO News*, *10*, 13.

National Center for Chronic Disease Prevention and Health Promotion (US) Office on Smoking and Health. (2014). *The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General*. Centers for Disease Control and Prevention (US).

Netzer, N. C., Stoohs, R. A., Netzer, C. M., Clark, K., & Strohl, K. P. (1999). Using the Berlin Questionnaire to Identify Patients at Risk for the Sleep Apnea Syndrome. *Annals of Internal Medicine*, *131*(7), 485–491. https://doi.org/10.7326/0003-4819-131-7-199910050-00002

Ng, A. (2018). *Machine Learning Yearning*. 1–118.

Office of the Surgeon General (US), Center for Mental Health Services (US), & National Institute of Mental Health (US). (2001). *Mental Health: Culture, Race, and Ethnicity*. Substance Abuse and Mental Health Services Administration (US).

Oliver, J. A., & McClernon, F. J. (2017). Precision Medicine in Addiction Research: Where Has the Time Gone?. *Addiction*, *112*(12), 2096–2097. https://doi.org/10.1111/add.14023

Partos, T. R., Borland, R., Yong, H.-H., Hyland, A., & Cummings, K. M. (2013). The Quitting Rollercoaster: How Recent Quitting History Affects Future Cessation Outcomes (Data From the International Tobacco Control 4-Country Cohort Study). *Nicotine & Tobacco Research*, *15*(9), 1578–1587. https://doi.org/10.1093/ntr/ntt025

Piper, Megan E., Cook, et al. (2017). Toward Precision Smoking Cessation Treatment II: Proximal Effects of Smoking Cessation Intervention Components on Putative Mechanisms of Action. *Drug and Alcohol Dependence*, *171*, 50–58. https://doi.org/10.1016/j.drugalcdep.2016.11.027

Piper, M. E., Fiore, M. C., Smith, S. S., Fraser, D., Bolt, D. M., Collins, L. M., Mermelstein, R., Schlam, T. R., Cook, J. W., Jorenby, D. E., Loh, W.-Y., & Baker, T. B. (2016). Identifying Effective Intervention Components for Smoking Cessation: A Factorial Screening Experiment. *Addiction (Abingdon, England)*, *111*(1), 129–141. https://doi.org/10.1111/add.13162

Piper, Megan E., Schlam, et al. (2017). Toward Precision Smoking Cessation Treatment I: Moderator Results from a Factorial Experiment. *Drug and Alcohol Dependence*, *171*, 59–65. https://doi.org/10.1016/j.drugalcdep.2016.11.025

Project Match Research Group. (1993). Project MATCH (Matching Alcoholism Treatment to Client Heterogeneity): Rationale and Methods for a Multisite Clinical Trial Matching Patients to Alcoholism Treatment. *Alcoholism, Clinical and Experimental Research*, *17*(6), 1130–1145. https://doi.org/10.1111/j.1530-0277.1993.tb05219.x

Project Match Research Group. (1998). Matching Alcoholism Treatments to Client Heterogeneity: Treatment Main Effects and Matching Effects on Drinking during Treatment. Project

MATCH Research Group. *Journal of Studies on Alcohol*, *59*(6), 631–639. https://doi.org/10.15288/jsa.1998.59.631

Rigotti, N. A., Kruse, G. R., Livingstone-Banks, J., & Hartmann-Boyce, J. (2022). Treatment of Tobacco Smoking: A Review. *JAMA*, *327*(6), 566–577. https://doi.org/10.1001/jama.2022.0395

Rosell, R., Carcereny, E., Gervais, R., Vergnenegre, A., Massuti, B., Felip, E., Palmero, R., Garcia-Gomez, R., Pallares, C., Sanchez, J. M., Porta, R., Cobo, M., Garrido, P., Longo, F., Moran, T., Insa, A., Marinis, F. D., Corre, R., Bover, I., … Paz-Ares, L. (2012). Erlotinib versus Standard Chemotherapy as First-Line Treatment for European Patients with Advanced EGFR Mutation-Positive Non-Small-Cell Lung Cancer (EURTAC): A Multicentre, Open-Label, Randomised Phase 3 Trial. *The Lancet Oncology*, *13*(3), 239–246. https://doi.org/10.1016/S1470-2045(11)70393-X

Roussidis, A., Kalkavoura, C., Dimelis, D., Theodorou, A., Ioannidou, I., Mellos, E., Mylonaki, T., Spyropoulou, A., & Yfantis, A. (2013). Reasons and Clinical Outcomes of Antipsychotic Treatment Switch in Outpatients with Schizophrenia in Real-Life Clinical Settings: The ETOS Observational Study. *Annals of General Psychiatry*, *12*, 42. https://doi.org/10.1186/1744-859X-12-42

RStudio Team. (2020). *RStudio: Integrated Development for R.*

Schlam, T. R., & Baker, T. B. (2013). Interventions for Tobacco Smoking. *Annual Review of Clinical Psychology*, *9*, 675–702. https://doi.org/10.1146/annurev-clinpsy-050212-185602

Schnoll, R. A., Patterson, F., Wileyto, E. P., Tyndale, R. F., Benowitz, N., & Lerman, C. (2009). Nicotine Metabolic Rate Predicts Successful Smoking Cessation with Transdermal Nicotine: A Validation Study. *Pharmacology Biochemistry and Behavior*, *92*(1), 6–11. https://doi.org/10.1016/j.pbb.2008.10.016

Shahab, L., Bauld, L., McNeill, A., & Tyndale, R. F. (2019). Does the Nicotine Metabolite Ratio Moderate Smoking Cessation Treatment Outcomes in Real-World Settings? A Prospective Study. *Addiction*, *114*(2), 304–314. https://doi.org/10.1111/add.14450

Siegel, S. D., Lerman, C., Flitter, A., & Schnoll, R. A. (2020). The Use of the Nicotine Metabolite Ratio as a Biomarker to Personalize Smoking Cessation Treatment: Current Evidence and Future Directions. *Cancer Prevention Research*, *13*(3), 261–272. https://doi.org/10.1158/1940-6207.CAPR-19-0259

Simons, J. S., & Gaher, R. M. (2005). The Distress Tolerance Scale: Development and Validation of a Self-Report Measure. *Motivation and Emotion*, *29*(2), 83–102. https://doi.org/10.1007/s11031-005-7955-3

Smets, E. M. A., Garssen, B., Bonke, B., & De Haes, J. C. J. M. (1995). The Multidimensional Fatigue Inventory (MFI) Psychometric Qualities of an Instrument to Assess Fatigue. *Journal of Psychosomatic Research*, *39*(3), 315–325. https://doi.org/10.1016/0022-3999(94)00125-O

Smith, S. S., Piper, M. E., Bolt, D. M., Fiore, M. C., Wetter, D. W., Cinciripini, P. M., & Baker, T. B. (2010). Development of the Brief Wisconsin Inventory of Smoking Dependence Motives. *Nicotine & Tobacco Research*, *12*(5), 489–499. https://doi.org/10.1093/ntr/ntq032

Smith, S. S., Piper, M. E., Bolt, D. M., Kaye, J. T., Fiore, M. C., & Baker, T. B. (2021). Revision of the Wisconsin Smoking Withdrawal Scale: Development of Brief and Long Forms. *Psychological Assessment*, *33*(3), 255–266. https://doi.org/10.1037/pas0000978

Snaith, R. P., Hamilton, M., Morley, S., Humayan, A., Hargreaves, D., & Trigwell, P. (1995). A Scale for the Assessment of Hedonic Tone the Snaith-Hamilton Pleasure Scale. *The British Journal of Psychiatry: The Journal of Mental Science*, *167*(1), 99–103. https://doi.org/10.1192/bjp.167.1.99

Soar, K., Dawkins, L., Robson, D., & Cox, S. (2020). Smoking amongst Adults Experiencing Homelessness: A Systematic Review of Prevalence Rates, Interventions and the Barriers and Facilitators to Quitting and Staying Quit. *Journal of Smoking Cessation*, *15*(2), 94–108. https://doi.org/10.1017/jsc.2020.11

Soyster, P. D., Ashlock, L., & Fisher, A. J. (2022). Pooled and Person-Specific Machine Learning Models for Predicting Future Alcohol Consumption, Craving, and Wanting to Drink: A Demonstration of Parallel Utility. *Psychology of Addictive Behaviors: Journal of the Society of Psychologists in Addictive Behaviors*, *36*(3), 296–306. https://doi.org/10.1037/adb0000666

Substance Abuse and Mental Health Services Administration. (2023). *Key Substance Use and Mental Health Indicators in the United States: Results from the 2022 National Survey on Drug Use and Health*.

Substance Abuse and Mental Health Services Administration (US), & Office of the Surgeon General (US). (2016). *Facing Addiction in America*. US Department of Health and Human Services.

Taylor, S., Zvolensky, M. J., Cox, B. J., Deacon, B., Heimberg, R. G., Ledley, D. R., Abramowitz, J. S., Holaway, R. M., Sandin, B., Stewart, S. H., Coles, M., Eng, W., Daly, E. S., Arrindell, W. A., Bouvard, M., & Cardenas, S. J. (2007). Robust Dimensions of Anxiety Sensitivity: Development and Initial Validation of the Anxiety Sensitivity Index-3. *Psychological Assessment*, *19*(2), 176–188. https://doi.org/10.1037/1040-3590.19.2.176

Taylor, V. R., & National Center for Chronic Disease Prevention and Health Promotion (U.S.). Division of Adult and Community Health. (2000). *Measuring Healthy Days: Population Assessment of Health-Related Quality of Life*.

Trautmann, S., Rehm, J., & Wittchen, H.-U. (2016). The Economic Costs of Mental Disorders. *EMBO Reports*, *17*(9), 1245–1249. https://doi.org/10.15252/embr.201642951

Tønnesen, P., Nørregaard, J., Säwe, U., & Simonsen, K. (1993). Recycling with Nicotine Patches in Smoking Cessation. *Addiction (Abingdon, England)*, *88*(4), 533–539. https://doi.org/10.1111/j.1360-0443.1993.tb02060.x

Wang, J., Simons-Morton, B. G., Farhat, T., Farhart, T., & Luk, J. W. (2009). Socio-Demographic Variability in Adolescent Substance Use: Mediation by Parents and Peers. *Prevention Science: The Official Journal of the Society for Prevention Research*, *10*(4), 387–396. https://doi.org/10.1007/s11121-009-0141-1

Webb, C. A., Trivedi, M. H., Cohen, Z. D., Dillon, D. G., Fournier, J. C., Goer, F., Fava, M., McGrath, P. J., Weissman, M., Parsey, R., Adams, P., Trombello, J. M., Cooper, C., Deldin, P., Oquendo, M. A., McInnis, M. G., Huys, Q., Bruder, G., Kurian, B. T., … Pizzagalli, D. A. (2019).

Personalized Prediction of Antidepressant v. Placebo Response: Evidence from the EMBARC Study. *Psychological Medicine*, *49*(7), 1118–1127. https://doi.org/10.1017/S0033291718001708

Weisz, J. R., Kuppens, S., Ng, M. Y., Vaughn-Coaxum, R. A., Ugueto, A. M., Eckshtain, D., & Corteselli, K. A. (2019). Are Psychotherapies for Young People Growing Stronger? Tracking Trends Over Time for Youth Anxiety, Depression, Attention-Deficit/Hyperactivity Disorder, and Conduct Problems. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *14*(2), 216–237. https://doi.org/10.1177/1745691618805436

Whiteford, H. A., Degenhardt, L., Rehm, J., Baxter, A. J., Ferrari, A. J., Erskine, H. E., Charlson, F. J., Norman, R. E., Flaxman, A. D., Johns, N., Burstein, R., Murray, C. J. L., & Vos, T. (2013). Global Burden of Disease Attributable to Mental and Substance Use Disorders: Findings from the Global Burden of Disease Study 2010. *Lancet (London, England)*, *382*(9904), 1575–1586. https://doi.org/10.1016/S0140-6736(13)61611-6

Wray, N. R., Lee, S. H., Mehta, D., Vinkhuyzen, A. A. E., Dudbridge, F., & Middeldorp, C. M. (2014). Research Review: Polygenic Methods and Their Application to Psychiatric Traits. *Journal of Child Psychology and Psychiatry, And Allied Disciplines*, *55*(10), 1068–1087. https://doi.org/10.1111/jcpp.12295

Wyant, K., Sant'Ana, S. J. K., Fronk, G., & Curtin, J. J. Machine Learning Models for Temporally Precise Lapse Prediction in Alcohol Use Disorder. *Psychopathology and Clinical Science*. https://doi.org/10.31234/osf.io/cgsf7

Youngstrom, E. A. (2014). A Primer on Receiver Operating Characteristic Analysis and Diagnostic Efficiency Statistics for Pediatric Psychology: We Are Ready to ROC. *Journal of Pediatric Psychology*, *39*(2), 204–221. https://doi.org/10.1093/jpepsy/jst062

Zheng, Y., Wiebe, R. P., Cleveland, H. H., Molenaar, P. C. M., & Harris, K. S. (2013). An Idiographic Examination of Day-to-Day Patterns of Substance Use Craving, Negative Affect, and Tobacco Use Among Young Adults in Recovery. *Multivariate Behavioral Research*, *48*(2), 241–266. https://doi.org/10.1080/00273171.2013.763012