Full title: Optimising supervised machine learning algorithms predicting cigarette cravings and lapses for a smoking cessation just-in-time adaptive intervention (JITAI)

Short title: Predicting high-risk moments in smoking cessation using machine learning

Corinna Leppin1, Jamie Brown1, Claire Garnett1,2, Dimitra Kale1, Tosan Okpako1, David Simons3, Olga Perski1,4,5

1 Department of Behavioural Science and Health, University College London, London, UK

2 School of Psychological Science, University of Bristol, Bristol, UK

3 Department of Anthropology, The Pennsylvania State University, University Park, PA, USA

4 Herbert Wertheim School of Public Health and Human Longevity Science, University of California, San Diego, CA, USA

5 Faculty of Social Sciences, Tampere University, Tampere, Finland

*Corresponding author: Corinna Leppin, Department of Behavioural Science and Health, University College London, 1-19 Torrington Place, London WC1E 7HB, UK. corinna.leppin.21@ucl.ac.uk

Abstract

This study aimed to optimise the balance between participant burden and performance of algorithms predicting high-risk moments for a smoking cessation just-in-time adaptive intervention (JITAI) by systematically varying ecological momentary assessment (EMA) prompt frequency, predictor count, and training data requirements.

Thirty-seven participants completed 16 EMAs per day for the first 10 days of their smoking cessation attempt. They reported their mood, context, behaviour, cravings, and smoking lapses. Random forest algorithms were used to predict lapses and cravings. The impacts of EMA prompt frequency, predictor count (reduced using recursive feature elimination with cross-validation [RFE-CV]), and inclusion of the test participant's data in the training set on algorithm performance, as indicated by the F1-score (balance of precision and recall; median [interquartile range]) were examined using ordered logistic mixed models.

Performance across out-of-sample individuals ranged from excellent to poor. However, several general trends emerged. Reducing EMA frequency modestly but significantly decreased model performance (e.g., 16 EMAs: 0.843 [0.480-0.946], 5 EMAs: 0.791 [0.390-0.912], 3 EMAs: 0.729 [0.286-0.897], $p < .001$ - .037). Predictor reduction had a negligible impact on algorithm performance and led to the selection of similar predictors across algorithms (all predictors: 0.773 [0.384-0.918], with RFE-CV: 0.773 [0.286-0.922]; $p = .313$). Algorithms predicting lapses outperformed those predicting cravings (lapses: 0.821 [0.323-0.943], cravings: 0.736 [0.340-0.904]; $p = .011$). Algorithms without any of the participants' own data in the training set outperformed algorithms that included varying proportions of the participant's own data (none: 0.795 [0.521-0.923], 10%: 0.765 [0.300-0.914], 20%: 0.736 [0.235-0.914], 30%: 0.763 [0.203-0.923]; $p < .001$- .012).

Although reducing EMA prompt frequency modestly decreases algorithm performance in predicting high-risk moments, this may be an acceptable trade-off for implementing a smoking cessation JITAI. Algorithm performance is unaffected by omitting user-specific training data and using fewer predictors. However, variable performance across individuals may limit scalability.

Author summary

This study investigated how to make digital support tools that help people quit smoking both effective and user-friendly. The goal is to develop a tool that offers help exactly when people are most at risk of smoking again. In this study, 37 people trying to quit smoking answered short surveys 16 times a day for 10 days about their behaviour, cravings, mood, and surroundings. This data was used to train computer models to predict when someone was likely to crave a cigarette or have a lapse. The study tested how fewer daily surveys, fewer questions, or different prediction methods would affect how well the prediction models work. We found that reducing the number of daily surveys reduced accuracy a bit, but asking fewer questions had little effect. This suggests simpler tools could still work well. Surprisingly, models that were not trained on a person's own data performed as well as or better than personalised ones, making them more practical for real-world use. However, we also saw that the models worked much better for some individuals than others, which suggests these tools might not be equally effective for everyone.

Introduction

Smoking remains a key cause of morbidity and mortality in the UK [1,2]. Improved interventions are needed to increase smoking cessation rates and improve population health in line with government targets [3]. Traditional behavioural interventions often focus on general motivation and coping strategies, but may be insufficient to prevent relapse [4,5]. Many people who smoke fail to actively seek out support at critical moments [6,7], and ensuing lapses often lead to full relapses [8-11]. Just-in-time adaptive interventions (JITAIs) aim to address this gap by delivering timely, tailored support during high-risk moments [12,13].

JITAIs rely on frequent assessment of tailoring variables (i.e., factors that may precipitate or shape high-risk moments), actively via ecological momentary assessments (EMAs) or passively via sensors. Decision rules (i.e., algorithms that specify how to combine data from tailoring variables) are then used to determine opportune moments for intervention delivery [13]. While early evidence suggests smoking cessation JITAIs are feasible [14-19] and can be effective [18], there remain questions over how to optimally balance the two (potentially conflicting) requirements of JITAIs: collecting enough data on psychological and contextual tailoring variables to effectively detect or predict high-risk moments and receptivity and creating a monitoring regime that is not too burdensome for users. Balancing these two requirements can be challenging in practice, as, all else being equal, an algorithm's ability to detect or predict states of need and receptivity, decipher their nature, and provide appropriate support is usually (but not always, cf. [20]) improved by more data and more sophisticated computing techniques. However, the time and effort involved in data collection and the drain on device battery resulting from passive sensors and complex modelling can be a burden for users [21,22].

The minimum sampling frequency at which psychological and contextual tailoring variables need to be assessed to collect data able to predict or detect high-risk moments well enough to prevent smoking lapses and effectively support smoking cessation is unclear. Given the frequency of smoking behaviour and the rapidity at which predictor variables may provoke lapses, hourly or more frequent measurements may be necessary to accurately model smoking dynamics [23,24]. However, previous research indicates that frequent or lengthy EMAs may reduce compliance and lead to disengagement with the intervention [25,26]. Although intensive data collection can enhance algorithm performance, most users are unlikely to tolerate more than four to five surveys with around 10 items per day for a prolonged period of time [16,19,21,27-30]. Limiting EMA frequency and item count per EMA may therefore be key to making a JITAI acceptable and maintaining user engagement.

JITAIs that are adaptive from the first interaction by using group-level algorithms, or gradually personalised "warm start" hybrid-level algorithms, appear to align better with user preferences than JITAIs that require extensive pre-quit training before providing tailored support [31,32]. While some previous evidence suggests that hybrid-level algorithms may be better at predicting high-risk moments during smoking cessation [9,29,33], they also require iterative retraining of the algorithms, which increases computational demand (and, particularly if done device-side, drain device battery, and if done server-side, have privacy risks) and adds complexity to the implementation process. This can limit their feasibility and acceptability, so a careful trade-off with performance compared to group-level algorithms is needed.

The current study aimed to train and test a random forest algorithm for a future smoking cessation JITAI. For this purpose, we conducted secondary analyses of previously collected EMA data [34] to explore how reducing the sampling frequency and the number of variables and varying the proportion of the test participants' own data in the training set affects the algorithm's ability to predict high-risk moments, operationalised either as lapses or as instances of high cravings, during a smoking cessation attempt.

The specific research question was: How does the performance of random forest algorithms predicting high-risk moments among people during the first 10 days of their smoking cessation attempt differ according to the frequency of EMA prompts per day, the number of predictor variables, the proportion of the test participant's own data in the training set, and whether high-risk moments are operationalised as smoking lapses or cigarette cravings?

Methods

Design

This study comprised secondary analyses of previously collected EMA data from 37 people who smoke during the first 10 days of their quit attempt [34]. The sample size was based on a simulation informed by the available literature; for more details, see https://osf.io/ywqpv. Participants received 16 EMAs per day (one every waking hour) to assess their mood, pain level, activity, social and physical context, cigarette craving, caffeine, alcohol, and nicotine consumption, and lapse incidence.

Ethical approval was obtained from the UCL Research Ethics Committee (Project ID: 15297.004). Participants were asked for their informed consent to share their anonymised research data with other researchers via an open science platform. The protocol for this study was preregistered on the Open Science Framework prior to the start of analyses (https://osf.io/tnu72/).

Recruitment and participants

Participants were recruited between May 2022 and March 2023, mostly through paid adverts on Facebook and Instagram. Eligibility requirements included being at least 18 years of age, regularly smoking cigarettes, residing in or around London, owning a smartphone, and being willing to set a quit date within seven days of the initial study visit. Participants were remunerated a minimum of £20 for participation, regardless of their EMA completion rate. Participants received £50 if they completed at least 50% of EMAs on at least 50% of study days and were informed they would earn an additional £5 for each day their EMA compliance exceeded 70%, up to a maximum of £100. It was made clear that payment was not tied to their smoking behaviour.

Procedure

After completing an online screening survey to confirm eligibility, eligible participants were invited to attend an in-person baseline visit. At the baseline visit, they completed a baseline survey and downloaded and received instructions for how to use the m-Path app [35], through which they received the EMAs throughout the study period. Participants were asked to set a quit date within the next seven days and encouraged to set it for the following day to capitalise on their motivation to stop smoking. Participants also received free access to the premium version of an evidence-based smoking cessation app [36]. At the end of the baseline visit, the in-person follow-up visit was scheduled.

The 10-day study period began on each participant's selected quit date. During this time, participants received 16 EMA prompts per day (one every waking hour on the hour or half-hour) via the m-Path app (signal-contingent sampling), tailored to their usual waking hours. EMAs needed to be completed within 30 minutes of receipt. Participants were also instructed to log lapses as they occurred using an in-app button (event-contingent sampling). Each EMA took around one to three minutes to complete and contained the same 22-25 items (depending on whether participants chose to add additional items; these additional items are not considered in this analysis, but more details on them can be found in the previous study using these data [34]). At the follow-up visit, data from study-related apps were exported, the m-Path app was deactivated, and participants received payment for their participation. Participants who reported that they had successfully quit smoking were asked to provide expired carbon monoxide measures using a Bedfont iCOquit smokerlyzer (https://www.icoquit.com/) to verify their abstinence (cut-off set at ≤8ppm). Full details on recruitment and procedure are reported in the previous study [34].

Measures

Overall, this secondary analysis used data from 34 variables: 12 between-person variables collected at the baseline questionnaire, and 22 within-person variables collected at every EMA. Full questionnaires and variable lists cand be found in S1 Appendix.

The between-person variables comprised two numerical variables (cigarettes per day; age in years), four binary variables (gender [four levels were available, but all participants selected either "man" or "woman"]); post-16 education; ever-use of behavioural support for smoking reduction or cessation; ever-use of pharmacological support for smoking reduction or cessation, and six categorical variables (occupation type; ethnicity; past quit attempt; time to first cigarette; motivation to stop scale [37]; season of data collection). The within-person variables comprised 14 quasi-continuous variables measured on 11-point Likert scales (ten variables measuring affect informed by the circumplex model of affect [38-40]: sad, irritable, stressed, anxious, bored, calm, content, happy, excited, enthusiastic; bodily pain; cigarette cravings; momentary motivation not to smoke; quitting-related self-efficacy), four binary variables (past-hour caffeine consumption; past-hour alcohol consumption; past-hour nicotine use; past-hour lapse), and four categorical variables (social context; activity; location; cigarette availability). For training and evaluating the algorithms, categorical variables were recoded into dummy variables.

Analyses

Data pre-processing

We used the pre-processed dataset from the previous study [34]. In this pre-processed dataset, participants with insufficient EMA adherence (defined as completion of ≤60% of prompts) were excluded. Missing data in the remaining analytical sample were imputed using the univariate Kalman filter, which is recommended for univariate time series data imputation [41] and has been previously used in analysis of intensive longitudinal data on health behaviours [42]. As a forward filter that only uses past and current information and does not consider future states, it is also suitable for implementation within an eventual JITAI.

Model specification

This study used group- and hybrid-level random forest (RF) algorithms to predict (a) lapses by or (b) cravings (dichotomised; high ≥ 7 on an 11-point Likert scale from 0 to 10) at one EMA based on levels of the predictor variables at the previous EMA and the values of the baseline variables.

RF algorithms are a machine learning method that combines the results of many decision trees to make more accurate and reliable predictions [43]. A decision tree is a flowchart-like model that splits the data into branches based on values of the predictor variables, aiming to group similar outcomes together [43]. In an RF, each tree is built using a random sample of the data and a random subset of predictors, which helps reduce overfitting and improves the model's ability to generalise to new data [43]. For every outcome by number of EMAs by set of predictors combination, the number of candidate predictors at each split and the minimum number of observations in a terminal node was tuned at the group-level, while the number of trees in the forest was set at 500, in line with recommendations for training RF algorithms [44-46]. As is standard, the threshold value was set at 0.5. All analyses were conducted using R [47], and the models were trained and tested using the tidymodels framework [48]. The complete analysis syntax and data can be found on Github (https://github.com/OlgaPerski/EMASENS_ML/tree/main) and the Open Science Framework (https://osf.io/tnu72/).

Selection of algorithm types

As the aim of this study was to train and test an algorithm for use in a future smoking cessation JITAI, we tested a single algorithm type for pragmatic reasons. We selected an RF algorithm because it has modest computational demands, is well suited for feature selection, does not make assumptions about the shape of the relationship between predictors and outcomes, and can handle a mix of continuous and categorical predictors as well as complex interactions [43,49,50]. RF algorithms also outperformed other algorithm types considered in the previous study (XGBoost, Support Vector Machine, Penalised Logistic Regression) [34].

Number and timing of prompts

We fit algorithms based on the data from 3, 4, 5, 6, and all 16 prompts per day. With a view to JITAI implementation considerations, we used quasi-random prompts evenly spaced throughout the day. To do this, we randomly selected prompts from blocks within the participant's daily schedule. The number of prompts to be tested was based on previous literature [16,19,21,27-30,51]. The prompts chosen for each number of prompts per day were as follows:

- 3: randomly choose one of 1-5, 6-11, 12-16
- 4: randomly choose one of 1-4, 5-8, 9-12, 13-16
- 5: randomly choose one of 1-3, 4-6, 7-9, 10-13, 14-16
- 6: randomly choose one of 1-3, 4-6, 7-8, 9-11, 12-14, 15-16
- All 16 prompts

When selecting the reduced number of prompts, we also included all event-based EMA reports (i.e., when a participant reported smoking outside of a prompt), regardless of the number of daily prompts tested.

Outcome definitions

We defined lapses as a lapse event that occurred between a participant's current and the next randomly selected/event-based EMA. The EMA was counted as a lapse if any lapse event was observed between these prompts. For high cravings, we recorded whether a high craving event was reported at the next randomly selected prompt. Additionally, we calculated prior lapses (as a predictor) by checking if a lapse occurred between the previous and current selected prompts.

In line with the previous study [34] and to allow for the calculation of the performance metrics, only participants with at least one observation in each outcome class were included. When constructing hybrid-level algorithms, participants were additionally required to have at least one observation in each outcome class in the testing set.

Number of predictors

We used data-driven feature selection to reduce the number of predictors. Feature selection describes machine learning techniques that aim to reduce the number of input dimensions while preserving the algorithm's performance [52]. In smoking cessation, filter [53] and embedded [30] feature selection methods have been previously used to improve the parsimony of models predicting smoking urges and lapses. The feature selection technique we used is recursive feature elimination with cross-validation (RFE-CV). RFE is an embedded feature selection method that is well-suited in this case, as it performs well even in the presence of correlated predictors [54], can handle both continuous and categorical predictors together and structured features, and does not have the same bias towards high cardinality features as some other methods. RFE-CV selects an optimal number of features without the need for pre-specification [55] and has been shown to have fairly high stability [56]. RFE-CV has shown promise in the epidemiology and healthcare fields [57-59]. To compare algorithms with fewer predictors to those with the full set of 57 predictors, we compared the performance of algorithms using the optimal set of predictors to that of algorithms with the full set of predictors. In cases where the optimal set of predictors was the full set, we compared the performance of algorithms using the top 15 predictors as identified by RFE-CV with that of algorithms using the full set of predictors. The full list of predictors used for each algorithm is provided in S2 Appendix.

Proportion of own data in the training set

For this study, group-level algorithms and hybrid-level algorithms were used to predict lapses and high cravings, with the latter incorporating the initial day or days of the test participant’s own data, along with the data from the other participants, in the training set. Including a participant’s data in the training set means using a portion of their data to help the machine learning model learn patterns during the training phase, before generating predictions on the participant’s later (unseen) data, which is reserved for testing the model’s performance. With a view to how a JITAI would be implemented in a real-world setting and to retain adequate amounts of data in the testing set, the specific amount of the participant’s own data in the training set was varied as 10% (first day), 20% (first two days), and 30% (first three days).

Performance

To evaluate the performance of the algorithms, we calculated the F1-score and its accompanying 95% confidence interval (CI). The F1-score is a commonly used composite

metric for assessing the performance of binary classifiers (i.e., algorithms that categorise observations into two classes, such as lapses and non-lapses or high cravings and low/no cravings). It is defined as the harmonic mean of precision and recall, providing a single measure that balances the trade-off between these two components. Precision (also known as positive predictive value) is the proportion of predicted positive cases (e.g., predicted lapses/high cravings) that are actual positives, while recall (or sensitivity) is the proportion of actual positive cases that are correctly identified by the model. The F1-score is calculated as:

$$F1 - score = 2 \times \left(\frac{precision \times recall}{precision + recall}\right)$$

This score ranges from 0 to 1, with 1 indicating perfect precision and recall, and thus ideal classification performance. The harmonic mean penalises extreme values more than the arithmetic mean, meaning the F1-score is low if either precision or recall is low. This makes it particularly suitable for imbalanced datasets, such as the one used in this study, where negative cases (non-lapses or low/no cravings) outnumber positive cases (lapses or high cravings), but the primary concern is maximising recall (i.e., the true-positive rate) to ensure that the vast majority of high-risk situations are identified [60-62]. The correct classification of positive cases is prioritised, because in an eventual JITAI, false positives (support provided in low-risk situations) are unlikely to be harmful to individuals, while false negatives (missing a high-risk situation) may set the individual on a trajectory towards full relapse [8-11]. To further evaluate performance, we also used the area under the receiver operating characteristic curve (ROC-AUC; overall ability to distinguish between classes), sensitivity (true positive rate), and specificity (true negative rate).

To examine patterns in performance metrics, ordered cumulative logistic mixed-effects models (CLMMs) were fitted using the ordinal package in R [63]. Each model treated the performance metric (e.g., ROC-AUC, F1 score, sensitivity, specificity) as an ordinal outcome, with raw scores divided into 100 equally sized percentiles. Ordered cumulative logistic mixed effects models were chosen over linear mixed effect models because F1-score, sensitivity, and specificity were highly skewed or bimodal, violating the distributional assumptions of linear models. Models included random intercepts for each participant to account for within-subject clustering, and fixed effects for the number of prompts per day (EMA frequency), outcome type (lapses versus cravings), use of feature selection (yes/no), and the proportion of own data in the training set (none, 10%, 20%, 30%). Separate models were fitted for each performance metric across outcomes, and stratified by outcome (cravings or lapses), with outcome excluded as a fixed effect in the stratified models. For fixed effects with more than two levels, marginal effects were estimated using the emmeans package [64] to compute pairwise comparisons with Tukey's adjustment for multiple comparisons.

Changes from pre-registration

Two changes were made to the pre-registered protocol. The first change was the addition of group-level algorithms. Originally, we only planned to train and evaluate hybrid-level algorithms. However, after conducting the initial analyses, we realised that there were a substantial number of participants who only had lapses in the first day(s) of their quit attempt, and for whom, consequently, hybrid-level algorithms could not be constructed, and saw that the performance of hybrid-level algorithms appeared unrelated to the amount of the

participant's own data in the training set. Therefore, we chose to additionally train and evaluate group-level algorithms and to include the proportion of own data in the training set as its own research question rather than as a sensitivity analysis. The second change was related to the method of feature selection. Originally, we planned to only compare the optimal set of predictors as determined by the RFE-CV algorithms with the full set of predictors. However, for a substantial number of algorithm configurations, the optimal set of predictors was all predictors. As that result would not allow us to compare the performance of algorithms with fewer predictors with that with the full set of predictors, we chose to select the 15 most important predictors as indicated by the RFE-CV procedure if that was the case and compare the algorithm with only these predictors with that with the full set of predictors.

Additionally, the preregistration did not specify a primary performance metric, but we focused on the F1-score in the main results for clarity and brevity. This choice was motivated by the F1-score's aforementioned advantages and the fact that ROC-AUC results largely aligned with conclusions based on the F1-score. Additionally, the preregistration did not specify how patterns in performance metrics would be analysed. Given the large number of algorithm configurations (5 prompt frequencies × 2 outcomes × 2 feature selection levels × 4 levels of the test participant's own data in the training set, each applied to every out-of-sample individual), we used ordered CLMMs to summarise performance. This approach offers a more interpretable and statistically robust overview than could be achieved through visual inspection or comparisons of specific algorithm configurations alone.

Results

Sample description

The analytic sample comprises n = 37 adult people who smoke from in and around London during the first 10 days of their quit attempt. Table 1 outlines participant characteristics.

Table 1 Participant characteristics

	All participants	**Participants with at least one lapse**
Characteristic	**N = 37**[1]	**N = 25**[1]
Age	42.97 (14.45)	42.48 (13.80)
Gender		
Female	23 (62.2%)	15 (60.0%)
Male	14 (37.8%)	10 (40.0%)
Occupation		
Non-manual	17 (45.9%)	12 (48.0%)
Manual	9 (24.3%)	5 (20.0%)
Other (e.g., student, unemployed, retired)	11 (29.7%)	8 (32.0%)
Post-16 educational qualifications	33 (89.2%)	22 (88.0%)
Ethnicity		
Asian or Asian British (any Asian background)	3 (8.1%)	3 (12.0%)
Black, Black British, Caribbean or African (any Black, Black British or Caribbean background)	1 (2.7%)	0 (0.0%)
Mixed or multiple ethnic groups (e.g., White and Black African, White and Asian)	2 (5.4%)	2 (8.0%)
Other ethnic group (i.e., Arab)	1 (2.7%)	1 (4.0%)
White (any White background)	30 (81.1%)	19 (76.0%)
Cigarettes per day	13.08 (6.50)	13.48 (6.72)
Time to first cigarette		
Within 5 minutes	9 (24.3%)	7 (28.0%)
6-30 minutes	16 (43.2%)	10 (40.0%)
31-60 minutes	6 (16.2%)	4 (16.0%)
After 60 minutes	6 (16.2%)	4 (16.0%)
Motivation to stop		

Characteristic	All participants N = 37[1]	Participants with at least one lapse N = 25[1]
I don't want to stop smoking	0 (0.0%)	0 (0.0%)
I think I should stop smoking but don't really want to	2 (5.4%)	2 (8.0%)
I want to stop smoking but haven't thought about when	6 (16.2%)	5 (20.0%)
I really want to stop smoking but don't know when I will	7 (18.9%)	3 (12.0%)
I want to stop smoking and hope to soon	9 (24.3%)	5 (20.0%)
I really want to stop smoking and intend to in the next 3 months	2 (5.4%)	2 (8.0%)
I really want to stop smoking and intend to in the next month	11 (29.7%)	8 (32.0%)
Past-year quit attempt		
No, never	3 (8.1%)	2 (8.0%)
Yes, but not in the past year	16 (43.2%)	9 (36.0%)
Yes, in the past year	18 (48.6%)	14 (56.0%)
Ever use of pharmacological support (e.g., nicotine replacement therapy, varenicline, e-cigarettes)	30 (81.1%)	20 (80.0%)
Ever use of behavioural support (e.g., counselling, telephone support, website, app)	16 (43.2%)	10 (40.0%)
Smoked since quit date		
In the last 24hrs	16 (43.2%)	14 (56.0%)
1-6 days ago	4 (10.8%)	3 (12.0%)
1-3 weeks ago	5 (13.5%)	4 (16.0%)
Have not smoked since my planned quit date	12 (32.4%)	4 (16.0%)
% Completed EMAs	76.86 (9.27)	76.38 (9.23)

[1]Mean (standard deviation); n (%); percentages may not always add up to 100% due to rounding

The majority of participants were female (62.2%) and of any White ethnic background (81.1%). The mean age was 42.97 years (Standard deviation (SD) = 14.45). Prior to quitting, participants smoked an average of 13.08 cigarettes per day (SD = 6.50). At follow-up, 43.2% had smoked in the past 24 hours, 10.8% smoked 1-6 days ago, and 32.4% had not smoked since their planned quit date. The average adherence to the EMAs was 76.86% (SD = 9.27). Among participants with at least one lapse (n = 25; 67.6% of the entire analytics sample), demographic and behavioural characteristics were similar: 60.0% were female and 76.0% identified with any White ethnicity. Their mean age was 42.48 years (SD = 13.80), and they smoked an average of 13.48 cigarettes per day (SD = 6.72). Despite reporting lapses during their EMAs, 16.0% reported having remained abstinent since their planned quit date at the follow-up visit. The average EMA adherence was 76.38% (SD = 9.23).

EMA description

Participants in the full analytic sample completed a total of 5,964 EMAs. Table 2 summarises key descriptives across the different prompt frequencies (16, 6, 5, 4, and 3 prompts per day). A version of this table restricted to participants with at least one lapse is available in S3 Appendix.

Event-contingent EMAs represented a small fraction of the total, ranging from 0.7% (16 prompts per day) to 4.0% (3 prompts per day). Lapse rates varied by prompt frequency, from 6.9% (16 prompts per day) to 20.8% (3 prompts per day). Due to the difference in calculation, the proportion of high cravings was higher and more consistent, ranging from 35.8% (16 prompts per day) to 39.9% (3 prompts per day). The proportion of EMAs recording lapses and high cravings varied substantially across individuals.

Table 2 Description of the observations

Prompts per day	EMAs				Lapses				High cravings			
	Number			Proportion event contingent	Overall proportion	Across individuals			Overall proportion	Across individuals		
	Total	Event contingent	Signal contingent			Median proportion	25th percentile proportion	75th percentile proportion		Median proportion	25th percentile proportion	75th percentile proportion
16	5,964	44	5,920	0.7%	6.9%	1.2%	0.0%	7.5%	35.8%	29.6%	11.1%	52.8%
6	2,204	44	2,160	2.0%	12.9%	3.3%	0.0%	18.2%	37.3%	30.9%	15.3%	59.4%
5	1,742	42	1,700	2.4%	15.2%	3.0%	0.0%	26.0%	38.1%	29.5%	16.9%	53.7%
4	1,401	41	1,360	2.9%	17.0%	3.7%	0.0%	28.8%	40.1%	29.8%	17.5%	57.7%
3	1,063	43	1,020	4.0%	20.8%	6.6%	0.0%	32.3%	39.0%	31.2%	16.3%	55.6%

Algorithm performance

Overall, algorithm performance varied substantially across both algorithm specifications and individual participants. F1-scores across algorithm specifications exhibited substantial dispersion, with medians often falling within the acceptable range, but means and interquartile ranges (IQRs) reflecting considerable variation across participants. Some participants' median F1-scores were near-perfect (i.e., 0.987) while others were not computable/zero due to a lack of true positives across algorithms. A full summary of key performance measures by algorithm specifications can be found in S4 Appendix, and a full summary of key performance measures by participant can be found in S5 Appendix.

Research question 1: Prompt frequency

Reducing the number of prompts per day led to statistically significant but modest declines in F1-score. As shown in Table 3, algorithms using the full set of 16 prompts per day achieved the highest median F1-score (0.843; IQR: 0.480-0.946), while those using 4 prompts had the lowest median score (0.689; IQR: 0.182-0.894). *Table 4* shows the summary of the CLMM of the F1-score by algorithm configurations. Compared to the reference category of 16 prompts per day, the odds of achieving higher F1-scores were significantly lower for algorithms using 3 prompts (Odds ratio (OR) = 0.626, 95% CI: 0.501-0.784, $p < .001$), 4 prompts (OR = 0.614, 95% CI: 0.490-0.770, $p < .001$), 5 prompts (OR = 0.787, 95% CI: 0.629-0.986, $p = .037$), and 6 prompts (OR = 0.727, 95% CI: 0.582-0.908, $p = .005$). Pairwise comparisons with adjusted p-values confirmed that the performance of algorithms with 16 daily prompts significantly exceeded that of algorithms with 3, 4, and 6 daily prompts (adjusted $p < .05$), though the difference between 16 and 5 daily prompts was not significant (adjusted $p = .227$). There were no significant differences between any of the lower prompt frequency algorithms.

Table 3 Summary of F1-score by algorithm configurations

	Median	**Median 95% CI lower bound**	**Median 95% CI upper bound**	**Mean**	**Standard deviation**	**25th percentile**	**75th percentile**
Prompts per day							
16	0.843	0.786	0.893	0.688	0.327	0.480	0.946
6	0.783	0.671	0.871	0.631	0.344	0.373	0.918
5	0.791	0.667	0.895	0.628	0.352	0.390	0.912
4	0.689	0.521	0.821	0.561	0.366	0.182	0.894
3	0.729	0.539	0.872	0.581	0.354	0.286	0.897
Feature selection use							
No	0.773	0.636	0.880	0.631	0.341	0.384	0.918
Yes	0.773	0.632	0.871	0.607	0.361	0.286	0.922
Outcome							
Lapses	0.821	0.704	0.909	0.643	0.362	0.323	0.943
Craving	0.736	0.591	0.855	0.605	0.344	0.340	0.904
Proportion of own data in training set							
None	0.795	0.682	0.889	0.684	0.300	0.521	0.923
10%	0.765	0.638	0.870	0.612	0.352	0.300	0.914

20%	0.736	0.593	0.855	0.588	0.367	0.235	0.914
30%	0.763	0.600	0.883	0.582	0.378	0.203	0.923

Table 4 Ordered (cumulative) logistic mixed-effects model of F1-Score (across outcomes)

	Odds ratio	95% CI	Standard error	p-value
Prompts per day				
16	1	ref	ref	ref
6	0.727	0.582-0.908	0.113	0.005
5	0.787	0.629-0.986	0.115	0.037
4	0.614	0.490-0.770	0.115	<0.001
3	0.626	0.501-0.784	0.114	<0.001
Feature selection				
No	1	ref	ref	ref
Yes	0.930	0.808-1.071	0.072	0.313
Outcome				
Lapses	1	ref	ref	ref
Cravings	0.799	0.671-0.950	0.088	0.011
Proportion of own data in training set				
None	1	ref	ref	ref
10%	0.777	0.639-0.945	0.100	0.012
20%	0.685	0.562-0.835	0.101	<0.001
30%	0.709	0.581-0.865	0.102	0.001

Number of ntiles: 100

This pattern was consistent across both outcomes. See S6 Appendix for the CLMM results stratified by outcome, full pairwise comparisons, and proportions of algorithms below the prespecified acceptability threshold (F1-score < 0.7).

Research question 2: Feature selection

Table 4 shows that reducing the number of predictors did not have a significant impact on performance (OR = 0.930, 95% CI: 0.808-1.071, p = .313). As can be seen in Table 3, algorithms using the full predictor set had a median F1-score of 0.773 (IQR: 0.384-0.918), identical to the median of those using a reduced set after feature selection (IQR: 0.286-0.922). When stratifying the model by outcome type, no significant difference was apparent either. See S6 Appendix for CLMM results stratified by outcome and proportions of algorithms below the prespecified acceptability threshold (F1-score < 0.7).

The predictors retained after feature selection were a mix of between-subject (time-invariant) and within-subject (time-varying) predictors and were largely similar across algorithms (see S2 Appendix). Key between-subject predictors such as age, cigarettes per day, occupation type, and general motivation to stop smoking were selected across algorithm specifications. Similarly, some within-subject psychological and contextual predictors, such as craving, self-efficacy, momentary motivation, and cigarette availability, were selected for most feature-reduced algorithms. Additionally, many feature-reduced algorithms retained some psychological variables relating to mood states (e.g., irritable, stressed, sad, anxious and happy) or physical pain. Some contextual predictors, such as location, current activity, or social context, were not selected for any of the feature-reduced algorithms. While the algorithms without feature selection used 57 predictors, this was reduced to 8 to 15 predictors for algorithms that did use feature selection.

Research question 3: Outcomes

As shown in Table 4, algorithms predicting cravings had significantly lower odds of achieving higher F1-scores than those predicting lapses (OR = 0.799, 95% CI: 0.671-0.950, p = .011). Table 3 summarises F1-scores by outcome. For lapse prediction, the median F1-score was 0.821 (IQR: 0.323-0.943), while for cravings prediction, it was 0.736 (IQR: 0.340-0.904). See S6 Appendix for the proportion of algorithms below the prespecified acceptability threshold (F1-score < 0.7).

Research question 4: Proportion of own data

Algorithms trained without any participant-specific data had the highest median F1-score (0.795; IQR: 0.521-0.923), followed by the algorithms using 10% (0.765; IQR: 0.300-0.914), 30% (0.763; IQR: 0.203-0.923), and 20% (0.736; IQR: 0.235-0.914) of the participant's own data (Table 3). This decrease in performance was statistically significant. As can be seen in Table 4, using models trained with none of the test participant's own data as the reference, the odds of higher F1-scores decreased significantly for algorithms trained with 10% (OR = 0.777, 95% CI: 0.639-0.945, p = .012), 20% (OR = 0.685, 95% CI: 0.562-0.835, p < .001), and 30% (OR = 0.709, 95% CI: 0.581-0.865, p = .001) of the test participant's own data. Pairwise comparisons with adjusted p-values confirmed that F1-scores for algorithms using none of the test participant's data were significantly higher than those using 20% (p = .001) or 30% (p = .004) of their own data. The comparison with algorithms using 10% of the test participant's own data approached significance (p = .056), while comparisons among the three hybrid conditions were not significant (all adjusted *p* > .600).

Stratifying the model by outcome revealed slightly divergent patterns in significance, but not the scores themselves. See S6 Appendix for the CLMM results stratified by outcome, full pairwise comparisons, and proportions of algorithms below the prespecified acceptability threshold (F1-score < 0.7).

Supplementary metrics

Across all research questions, the pattern of performance indicated by the supplementary metrics (ROC-AUC, sensitivity, and specificity) largely mirrored the trends in F1-score, but there were some differences, particularly in terms of the statistical significance of the patterns. ROC-AUC remained moderately high and was less impacted by the instances of zero-true-positive cases that depressed F1-scores. Sensitivity was strong across conditions, while specificity stayed lower and more stable, helping explain why some algorithms with poor F1-scores still had acceptable ROC-AUCs. Lower prompt frequencies resulted in statistically significant reductions in F1-score and ROC-AUC, but not in sensitivity or specificity. Algorithms predicting lapses significantly outperformed those predicting cravings across all metrics. Feature selection had no overall significant impact on overall performance metrics, but did lead to significant improvements in the specificity of lapse prediction algorithms. Including participants' own data harmed F1-score and sensitivity, while leaving specificity unaffected and only marginally affecting ROC-AUC.

See S7 Appendix (ROC-AUC), S8 Appendix (sensitivity), S9 Appendix (specificity) for metric summaries, CLMM results, pairwise comparisons, and proportions of algorithms below the prespecified acceptability threshold (F1-score and sensitivity < 0.7; ROC-AUC and specificity < 0.5).

Discussion

Summary of main findings

This study adds to the growing literature demonstrating that it is possible to predict smoking behaviour from EMA data using machine learning algorithms [34,53,65,66]. It does so by systematically examining the effects of prompt frequency, feature selection, prediction target (lapse versus craving), and use of participant-specific data in training on the performance of RF machine learning algorithms to underpin a future smoking cessation JITAI. Although the performance of the algorithms varied substantially across individuals, some general trends were observed. Algorithm performance measured using the F1-score (and ROC-AUC) declined modestly but significantly with reduced prompt frequency. Feature reduction via RFE did not significantly impact performance and led to the selection of similar predictors across algorithm specifications. Prediction of lapses outperformed prediction of cravings. Unexpectedly, there was an indication that group-level algorithms, which did not include any of the participants' own data in the training set, outperformed hybrid-level algorithms.

Practical implications

These findings suggest that it may be feasible to implement algorithms predicting high-risk moments, particularly lapses, within a smoking cessation JITAI without imposing an excessive burden on users. The somewhat greater difficulty of predicting cravings may be due to the algorithms attempting to predict high cravings at the next EMA and lapses by the next EMA. As a fluctuating measure, the former may simply be inherently more difficult to predict, especially with longer and more variable gaps between EMAs. Although this study showed that performance declines with less data density, algorithms with five or six prompts per day still generally performed in the acceptable range.

Consistent with the nature of RF models, which intrinsically prioritise the most relevant features, our results did not show a statistically significant impact of an additional, explicit feature selection step that reduces the number of predictor variables. Although we cannot conclude a true absence of an effect, the point estimate of the difference was small. Additionally, the retained predictors were largely consistent across algorithms, indicating a fairly robust relationship between these predictors and smoking lapses and cravings. Given the substantial benefit of reducing survey burden, using a more parsimonious model with fewer psychological and contextual tailoring variables appears to be a justifiable and practical strategy. Nonetheless, concerns about generalisability remain, discussed further below. Previous research indicates that users of a smoking cessation JITAI may not be willing to respond to more than four to five surveys with around 10 items per day [16,19,21,27-30]. Therefore, using one of the lapse risk algorithms using five prompts/day and feature selection may enable researchers to develop a smoking cessation JITAI that has the potential to be both acceptable and effective.

Importantly, despite substantial differences in how well the algorithms performed for different individuals, algorithms trained without participant-specific data performed as well as or better than algorithms incorporating small proportions of the test participant's own data. This has favourable implications in terms of feasibility as it suggests that individualised algorithm training is not necessary to maintain acceptable performance. However, there may still be limitations in terms of generalisability (discussed further below), as the algorithms are trained on individuals who have at least one instance of lapse or high craving and may therefore work less well on those who have none or very few of these events. With these caveats in

mind, a group-level lapse-risk algorithm using five prompts per day and feature selection may nevertheless be the best way going forward for a future JITAI based on this research.

However, the variable algorithm performance across individuals, with performance for some individuals being consistently very low, raises concerns with a view to implementation, as it means that the algorithms trained and tested in this study will likely not be effective for a subset of potential users. More research is needed to identify why this is the case and whether and how algorithms with better predictive performance can be constructed, by, e.g., more explicitly accounting for the multilevel nature of the data.

Links to previous research

The predictors most consistently selected through RFE in this study, namely cravings, self-efficacy, momentary motivation, cigarette availability, and affective states, are largely consistent with factors identified as important psychological and contextual variables in prior research identifying predictors of smoking lapses [30,34,66-71]. Location, current activity, and social context were not selected for any feature-reduced algorithms in this study. This aligns with mixed evidence regarding the predictive utility of these contextual features; previous research has found variables relating to these concepts to be key predictors in some models, but not others [34,53,66,67,69,70]. A potential reason why these variables were not selected for any of the feature-reduced algorithms in this study may be how they were operationalised (the variables in this study were fairly fine-grained) and the fact that feature selection was done at the group-level (these variables may still be highly relevant for some individuals). Overall, this suggests that while there is some consistency in the key psychological and contextual predictors of lapse risk, the relevance of certain factors, such as location and social setting, may be more variable, potentially depending on the modelling approach, operationalisations, the dataset, and/or differences between individuals or populations.

The finding that group-level algorithms slightly outperformed hybrid-level algorithms contrasts with results from previous analyses of machine learning algorithms to predict smoking lapses [66], including the previous study using the same dataset [34]. A likely explanation for this discrepancy is the difference in how the sub-sample of participants' own data that was included in the training set for the hybrid-level algorithms was selected. In the previous studies, hybrid-level algorithms were trained with data that included a randomly selected 20% or 40% of each participant's data. In the present study, to simulate how a hybrid-level algorithm would operate if implemented as part of a real-world smoking cessation intervention, we used the first 10%, 20%, or 30% of each participant's data chronologically. This difference in data selection may have affected the representativeness and diversity of the training data, particularly if, as research using time-varying effect models suggest [67,68,72-77], the predictors of smoking lapses and cravings early on a quit attempt are different than the predictors at later points.

Strengths, limitations, and directions for future research

This study has several notable strengths. It is, to the authors' knowledge, the first study to assess how EMA sampling frequency, predictor count, and training data composition affect algorithm performance for predicting smoking lapses and cravings. These factors are critical considerations when developing a JITAI, and by systematically examining them, this study can provide a reference for making informed decisions about the necessary trade-offs. Additionally, the use of high-frequency EMA data over a critical 10-day period provides detailed insight into momentary risk factors. Furthermore, the testing of a large number of algorithm configurations and cross-participant validation allows for a nuanced assessment of

generalisability and individual-level variability in performance. Finally, the preregistered protocol and open sharing of data and code promote transparency and reproducibility.

There are also several ways for future research to extend and build on the findings from this study. First, as a secondary analysis of previously collected data, this study used a randomly selected subset of EMAs to explore the impact of reducing sampling frequency. Future studies could assess the ecological validity of the results by experimentally varying prompt frequencies to explore how participants respond to less intensive monitoring regimes and assess the real-world impact of different prompt frequencies on algorithm performance. Second, our analyses focused on short-term prediction, using data from the immediately preceding EMA. Future work could explore models that account for longer-term behavioural patterns or temporal dynamics across the quit attempt to explore whether that would improve predictive performance. Third, the dataset covers only the first 10 days of participants' smoking cessation attempts. While this period is critical due to the high risk of lapse and relapse, predictors of lapses may change over time [67,68,72-77]. Future research could extend the time period under investigation.

Although algorithm performance varied substantially across individuals, investigating why some participants' smoking behaviour was harder to predict than others is beyond the scope of this study. Future research could aim to identify person-level characteristics that influence algorithm performance to inform the development of more personalised or stratified approaches. Additionally, as the sample in this study was self-selected, geographically localised, and skewed toward participants willing to engage in intensive EMA protocols, future studies should assess algorithm performance in more diverse and representative populations. This is particularly important given evidence that smoking patterns and lapse predictors may vary by socioeconomic status [78,79], and given that occupation type emerged as a key predictor in this study (see S2 Appendix). While the feature selection procedure identified a relatively stable set of predictors across models, previous work has shown that feature importance can vary substantially across contexts [30,52]. Further work is needed to evaluate the generalisability of these predictors and to better understand their interpretability, including the direction and strength of their association with outcomes.

Finally, the algorithms in this study employed a standard decision threshold of 0.5 to classify predicted probabilities when distinguishing between lapses/non-lapses and high cravings/no or low cravings. While intuitively sensible to get accurate predictions and appropriate for exploratory purposes, future work focused on clinical or real-world implementation may benefit from adjusting this threshold to optimise either sensitivity or specificity.

Conclusion

Overall, the findings of this study suggest that it is possible to predict high-risk moments, particularly if defined as lapses, during a smoking cessation attempt from EMA data using a limited number of time-varying predictors without imposing excessive user or computational burden. Notably, group-level algorithms without participant-specific data often performed as well as or better than hybrid models, suggesting that personalised algorithm training is not essential for effective prediction. Although performance declines modestly with reduced prompt frequency, this is a trade-off that may be acceptable in the real-world implementation of JITAIs. However, considerable variability in algorithm performance across individuals raises concerns about generalisability and scalability and highlights the need for more research into who high-risk prediction algorithms work for and why. Together, these findings provide valuable guidance for the pragmatic development of digital interventions that aim to deliver support at high-risk moments while maintaining user acceptability and feasibility.

References

1. NHS Digital. Statistics on Public Health, England 2023. 2024. Available: https://digital.nhs.uk/data-and-information/publications/statistical/statistics-on-public-health/2023

2. Reitsma MB, Fullman N, Ng M, Salama JS, Abajobir A, Abate KH, et al. Smoking prevalence and attributable disease burden in 195 countries and territories, 1990-2015: a systematic analysis from the Global Burden of Disease Study 2015. The Lancet. 2017;389: 1885-1906. doi:10.1016/S0140-6736(17)30819-X

3. Song F, Elwell-Sutton T, Naughton F, Gentry S. Future smoking prevalence by socioeconomic status in England: a computational modelling study. Tob Control. 2021;30: 380. doi:10.1136/tobaccocontrol-2019-055490

4. Hartmann-Boyce J, Ordóñez-Mena JM, Livingstone-Banks J, Fanshawe TR, Lindson N, Freeman SC, et al. Behavioural programmes for cigarette smoking cessation: investigating interactions between behavioural, motivational and delivery components in a systematic review and component network meta-analysis. Addiction. 2022;117: 2145-2156. doi:10.1111/add.15791

5. Livingstone-Banks J, Norris E, Hartmann-Boyce J, West R, Jarvis M, Hajek P. Relapse prevention interventions for smoking cessation. Cochrane Database Syst Rev. 2019. doi:10.1002/14651858.CD003999.pub5

6. Brendryen H, Kraft P. Happy Ending: a randomized controlled trial of a digital multi-media smoking cessation intervention. Addiction. 2008;103: 478-484. doi:10.1111/j.1360-0443.2007.02119.x

7. Ubhi HK, Michie S, Kotz D, Wong WC, West R. A Mobile App to Aid Smoking Cessation: Preliminary Evaluation of SmokeFree28. J Med Internet Res. 2015;17: e3479. doi:10.2196/jmir.3479

8. Ashare RL, Wileyto EP, Perkins KA, Schnoll RA. The first seven days of a quit attempt predicts relapse: Validation of a measure for screening medications for nicotine dependence. J Addict Med. 2013;7: 249.

9. Deiches JF, Baker TB, Lanza S, Piper ME. Early Lapses in a Cessation Attempt: Lapse Contexts, Cessation Success, and Predictors of Early Lapse. Nicotine Tob Res. 2013;15: 1883-1891. doi:10.1093/ntr/ntt074

10. Kenford SL, Fiore MC, Jorenby DE, Smith SS, Wetter D, Baker TB. Predicting smoking cessation: Who will quit with and without the nicotine patch. Jama. 1994;271: 589-594.

11. Shadel WG, Martino SC, Setodji C, Cervone D, Witkiewitz K, Beckjord EB, et al. Lapse-induced surges in craving influence relapse in adult smokers: an experimental investigation. Health Psychol. 2011;30: 588.

12. Hardeman W, Houghton J, Lane K, Jones A, Naughton F. A systematic review of just-in-time adaptive interventions (JITAIs) to promote physical activity. Int J Behav Nutr Phys Act. 2019;16: 31. doi:10.1186/s12966-019-0792-7

13. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design

Principles for Ongoing Health Behavior Support. Ann Behav Med. 2018;52: 446-462. doi:10.1007/s12160-016-9830-8

14. Battalio SL, Conroy DE, Dempsey W, Liao P, Menictas M, Murphy S, et al. Sense2Stop: a micro-randomized trial using wearable sensors to optimize a just-in-time-adaptive stress management intervention for smoking relapse prevention. Contemp Clin Trials. 2021;109: 106534. doi:10.1016/j.cct.2021.106534

15. Businelle MS, Ma P, Kendzor DE, Frank SG, Vidrine DJ, Wetter DW. An ecological momentary intervention for smoking cessation: evaluation of feasibility and effectiveness. J Med Internet Res. 2016;18.

16. Hébert ET, Ra CK, Alexander AC, Helt A, Moisiuc R, Kendzor DE, et al. A mobile just-in-time adaptive intervention for smoking cessation: pilot randomized controlled trial. J Med Internet Res. 2020;22: e16907. doi:10.2196/16907

17. Naughton F, Hopewell S, Lathia N, Schalbroeck R, Brown C, Mascolo C, et al. A Context-Sensing Mobile Phone App (Q Sense) for Smoking Cessation: A Mixed-Methods Study. JMIR Mhealth Uhealth. 2016;4: e106. doi:10.2196/mhealth.5787

18. Naughton F, Hope A, Siegele-Brown C, Grant K, Barton G, Notley C, et al. An automated, online feasibility randomized controlled trial of a just-in-time adaptive intervention for smoking cessation (Quit Sense). Nicotine Tob Res. 2023;25: 1319-1329. doi:10.1093/ntr/ntad032

19. Yang M-J, Sutton SK, Hernandez LM, Jones SR, Wetter DW, Kumar S, et al. A Just-In-Time Adaptive intervention (JITAI) for smoking cessation: Feasibility and acceptability findings. Addict Behav. 2023;136: 107467. doi:10.1016/j.addbeh.2022.107467

20. Meng X-L, Xie X. I Got More Data, My Model is More Refined, but My Estimator is Getting Worse! Am I Just Dumb? Econom Rev. 2014;33: 218-250. doi:10.1080/07474938.2013.808567

21. Burke LE, Shiffman S, Music E, Styn MA, Kriska A, Smailagic A, et al. Ecological Momentary Assessment in Behavioral Research: Addressing Technological and Human Participant Challenges. J Med Internet Res. 2017;19: e77. doi:10.2196/jmir.7138

22. Materia FT, Smyth JM. Acceptability of Intervention Design Factors in mHealth Intervention Research: Experimental Factorial Study. JMIR Mhealth Uhealth. 2021;9: e23303. doi:10.2196/23303

23. Chevance G, Perski O, Hekler EB. Innovative methods for observing and changing complex health behaviors: four propositions. Transl Behav Med. 2021;11: 676-685. doi:10.1093/tbm/ibaa026

24. Ferguson SG, Shiffman S. The relevance and treatment of cue-induced cravings in tobacco dependence. J Subst Abuse Treat. 2009;36: 235-243. doi:10.1016/j.jsat.2008.06.005

25. Dao KP, De Cocker K, Tong HL, Kocaballi AB, Chow C, Laranjo L. Smartphone-delivered ecological momentary interventions based on ecological momentary assessments to promote health behaviors: systematic review and adapted checklist for reporting ecological momentary assessment and intervention studies. JMIR MHealth UHealth. 2021;9: e22890. doi:10.2196/22890

26. Kabir KS, Kenfield SA, Van Blarigan EL, Chan JM, Wiese J. Ask the Users: A Case Study of Leveraging User-Centered Design for Designing Just-in-Time Adaptive Interventions (JITAIs). Proc ACM Interact Mob Wearable Ubiquitous Technol. 2022;6: 1-21.

27. Crochiere RJ, Zhang F (Zoe), Juarascio AS, Goldstein SP, Thomas JG, Forman EM. Comparing ecological momentary assessment to sensor-based approaches in predicting dietary lapse. Transl Behav Med. 2021;11: 2099-2109. doi:10.1093/tbm/ibab123

28. Forman EM, Goldstein SP, Zhang F, Evans BC, Manasse SM, Butryn ML, et al. OnTrack: development and feasibility of a smartphone app designed to predict and prevent dietary lapses. Transl Behav Med. 2019;9: 236-245. doi:10.1093/tbm/iby016

29. Goldstein SP, Zhang F, Thomas JG, Butryn ML, Herbert JD, Forman EM. Application of Machine Learning to Predict Dietary Lapses During Weight Loss. J Diabetes Sci Technol. 2018;12: 1045-1052. doi:10.1177/1932296818775757

30. Hébert ET, Suchting R, Ra CK, Alexander AC, Kendzor DE, Vidrine DJ, et al. Predicting the first smoking lapse during a quit attempt: A machine learning approach. Drug Alcohol Depend. 2021;218: 108340. doi:10.1016/j.drugalcdep.2020.108340

31. Ubhi HK, Kotz D, Michie S, van Schayck OCP, West R. A comparison of the characteristics of iOS and Android users of a smoking cessation app. Transl Behav Med. 2017;7: 166-171. doi:10.1007/s13142-016-0455-z

32. Cobb CO, Niaura RS, Donaldson EA, Graham AL. Quit Now? Quit Soon? Quit When You’re Ready? Insights About Target Quit Dates for Smoking Cessation From an Online Quit Date Tool. J Med Internet Res. 2014;16: e55. doi:10.2196/jmir.3086

33. Shiffman S, Hickcox M, Paty JA, Gnys M, Richards T, Kassel JD. Individual differences in the context of smoking lapse episodes. Smok Cessat Clin Res Dir. 1997;22: 797-811. doi:10.1016/S0306-4603(97)00063-4

34. Perski O, Kale D, Leppin C, Okpako T, Simons D, Goldstein SP, et al. Supervised machine learning to predict smoking lapses from Ecological Momentary Assessments and sensor data: Implications for just-in-time adaptive intervention development. PLOS Digit Health. 2024;3: e0000594. doi:10.1371/journal.pdig.0000594

35. Mestdagh M, Verdonck S, Piot M, Niemeijer K, Tuerlinckx F, Kuppens P, et al. m-Path: An easy-to-use and flexible platform for ecological momentary assessment and intervention in behavioral research and clinical practice. PsyArXiv Prepr. 2022.

36. Jackson SE, Perski O, Crane D, Michie S, West R, Brown J. Effectiveness of an offer of the Smoke Free smartphone application for smoking cessation: protocol for a randomized controlled trial. Addiction. 2019;114: 2078-2086. doi:10.1111/add.14652

37. Kotz D, Brown J, West R. Predictive validity of the Motivation To Stop Scale (MTSS): A single-item measure of motivation to stop smoking. Drug Alcohol Depend. 2013;128: 15-19. doi:10.1016/j.drugalcdep.2012.07.012

38. Barrett LF, Gross J, Christensen TC, Benvenuto M. Knowing what you’re feeling and knowing what to do about it: Mapping the relation between emotion differentiation and emotion regulation. Cogn Emot. 2001;15: 713-724. doi:10.1080/02699930143000239

39. Posner J, Russell JA, Peterson BS. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev Psychopathol. 2005/11/01 ed. 2005;17: 715-734. doi:10.1017/S0954579405050340

40. Russell JA. A circumplex model of affect. J Pers Soc Psychol. 1980;39: 1161. doi:10.1037/h0077714

41. Moritz S, Sardá A, Bartz-Beielstein T, Zaefferer M, Stork J. Comparison of different methods for univariate time series imputation in R. ArXiv Prepr ArXiv151003924. 2015.

42. Chevance G, Baretta D, Heino M, Perski O, Olthof M, Klasnja P, et al. Characterizing and predicting person-specific, day-to-day, fluctuations in walking behavior. PLoS One. 2021;16: e0251659. doi:10.1371/journal.pone.0251659

43. Rhys H. Machine Learning with R, the tidyverse, and mlr. 1st edition. New York: Manning; 2020.

44. Probst P, Wright MN, Boulesteix A-L. Hyperparameters and tuning strategies for random forest. WIREs Data Min Knowl Discov. 2019;9: e1301. doi:10.1002/widm.1301

45. Probst P, Boulesteix A-L. To tune or not to tune the number of trees in random forest. J Mach Learn Res. 2018;18: 1-18.

46. Scornet E. Tuning parameters in random forests. ESAIM Proc Surv. 2017;60: 144-162. doi:10.1051/proc/201760144

47. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2024. Available: https://www.R-project.org/

48. Kuhn M, Wickham H. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. 2020. Available: https://www.tidymodels.org

49. Singh KJ, Meitei AJ, Alee NT, Kriina M, Haobijam NS. Machine learning algorithms for predicting smokeless tobacco status among women in Northeastern States, India. Int J Syst Assur Eng Manag. 2022;13: 2629-2639. doi:10.1007/s13198-022-01720-3

50. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol Methods. 2009;14: 323. doi:10.1037/a0016973

51. Leppin C, Okpako T, Garnett C, Perski O, Brown J. Exploring perspectives on digital smoking cessation just-in-time adaptive interventions: A focus group study with adult smokers and smoking cessation professionals. PLOS Digit Health. 2025;4: e0000705. doi:10.1371/journal.pdig.0000705

52. Tang J, Alelyani S, Liu H. Feature selection for classification: A review. Data Classif Algorithms Appl. 2014; 37.

53. Dumortier A, Beckjord E, Shiffman S, Sejdić E. Classifying smoking urges via machine learning. Comput Methods Programs Biomed. 2016;137: 203-213. doi:10.1016/j.cmpb.2016.09.016

54. Gregorutti B, Michel B, Saint-Pierre P. Correlation and variable importance in random forests. Stat Comput. 2017;27: 659-678. doi:10.1007/s11222-016-9646-1

55. Cheng Y-C, Chang H-T, Lin C-Y, Chang H-Y. Predicting Credit Risk in Peer-to-Peer Lending: A Machine Learning Approach with Few Features. 2021 International Conference on Technologies and Applications of Artificial Intelligence (TAAI). 2021. pp. 295-300. doi:10.1109/TAAI54685.2021.00064

56. Akkaya B. The Effect of Recursive Feature Elimination with Cross-Validation Method on Classification Performance with Different Sizes of Datasets. 2021. pp. 4-6.

57. Agaal A, Essgaer M. Influence of Feature Selection Methods on Breast Cancer Early Prediction Phase using Classification and Regression Tree. 2022 International Conference on Engineering & MIS (ICEMIS). 2022. pp. 1-6. doi:10.1109/ICEMIS56295.2022.9914078

58. Choi SG, Oh M, Park D, Lee B, Lee Y, Jee SH, et al. Comparisons of the prediction models for undiagnosed diabetes between machine learning versus traditional statistical methods. Sci Rep. 2023;13: 13101. doi:10.1038/s41598-023-40170-0

59. Habenicht R, Fehrmann E, Blohm P, Ebenbichler G, Fischer-Grote L, Kollmitzer J, et al. Machine Learning Based Linking of Patient Reported Outcome Measures to WHO International Classification of Functioning, Disability, and Health Activity/Participation Categories. J Clin Med. 2023;12. doi:10.3390/jcm12175609

60. Diallo R, Edalo C, Awe OO. Machine Learning Evaluation of Imbalanced Health Data: A Comparative Analysis of Balanced Accuracy, MCC, and F1 Score. In: Awe OO, A. Vance E, editors. Practical Statistical Learning and Data Science Methods: Case Studies from LISA 2020 Global Network, USA. Cham: Springer Nature Switzerland; 2025. pp. 283-312. doi:10.1007/978-3-031-72215-8_12

61. Parker C. An Analysis of Performance Measures for Binary Classifiers. 2011 IEEE 11th International Conference on Data Mining. 2011. pp. 517-526. doi:10.1109/ICDM.2011.21

62. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag. 2009;45: 427-437. doi:10.1016/j.ipm.2009.03.002

63. Christensen RHB. ordinal—Regression Models for Ordinal Data. 2023. Available: https://CRAN.R-project.org/package = ordinal

64. Lenth RV. emmeans: Estimated Marginal Means, aka Least-Squares Means. 2025. Available: https://rvlenth.github.io/emmeans/

65. Abo-Tabik M, Costen N, Darby J, Benn Y. Decision Tree Model of Smoking Behaviour. 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI 2019, Leicester, United Kingdom, August 19-23, 2019. IEEE; 2019. pp. 1746-1753. doi:10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00311

66. Perski O, Li K, Pontikos N, Simons D, Goldstein SP, Naughton F, et al. Classification of Lapses in Smokers Attempting to Stop: A Supervised Machine Learning Approach Using Data From a Popular Smoking Cessation Smartphone App. Nicotine Tob Res. 2023;25: 1330-1339. doi:10.1093/ntr/ntad051

67. Koslovsky MD, Hébert ET, Businelle MS, Vannucci M. A Bayesian time-varying effect model for behavioral mHealth data. Ann Appl Stat. 2020;14: 1878-1902. doi:10.1214/20-AOAS1402

68. Koslovsky MD, Hébert ET, Swartz MD, Chan W, Leon-Novelo L, Wilkinson AV, et al. The Time-Varying Relations Between Risk Factors and Smoking Before and After a Quit Attempt. Nicotine Tob Res. 2018;20: 1231-1236. doi:10.1093/ntr/ntx225

69. Koslovsky MD, Swartz MD, Chan W, Leon-Novelo L, Wilkinson AV, Kendzor DE, et al. Bayesian Variable Selection for Multistate Markov Models with Interval-censored Data in an Ecological Momentary Assessment Study of Smoking Cessation. Biometrics. 2018;74: 636-644. doi:10.1111/biom.12792

70. Suchting R, Hébert ET, Ma P, Kendzor DE, Businelle MS. Using Elastic Net Penalized Cox Proportional Hazards Regression to Identify Predictors of Imminent Smoking Lapse. Nicotine Tob Res. 2019;21: 173-179. doi:10.1093/ntr/ntx201

71. Gwaltney CJ, Shiffman S, Balabanis MH, Paty JA. Dynamic self-efficacy and outcome expectancies: prediction of smoking lapse and relapse. J Abnorm Psychol. 2005;114: 661. doi:10.1037/0021-843X.114.4.661

72. Dermody SS, Shiffman S. The time-varying effect of alcohol use on cigarette smoking relapse risk. Addict Behav. 2020;102: 106192. doi:10.1016/j.addbeh.2019.106192

73. Lanza ST, Vasilenko SA, Liu X, Li R, Piper ME. Advancing the Understanding of Craving During Smoking Cessation Attempts: A Demonstration of the Time-Varying Effect Model. Nicotine Tob Res. 2014;16: S127-S134. doi:10.1093/ntr/ntt128

74. Potter LN, Haaland BA, Lam CY, Cambron C, Schlechter CR, Cinciripini PM, et al. A time-varying model of the dynamics of smoking lapse. Health Psychol. 2021;40: 40. doi:10.1037/hea0001036

75. Shiyko M, Naab P, Shiffman S, Li R. Modeling Complexity of EMA Data: Time-Varying Lagged Effects of Negative Affect on Smoking Urges for Subgroups of Nicotine Addiction. Nicotine Tob Res. 2014;16: S144-S150. doi:10.1093/ntr/ntt109

76. Shiyko M, Lanza ST, Tan X, Li R, Shiffman S. Using the Time-Varying Effect Model (TVEM) to Examine Dynamic Associations between Negative Affect and Self Confidence on Smoking Urges: Differences between Successful Quitters and Relapsers. Prev Sci. 2012;13: 288-299. doi:10.1007/s11121-011-0264-z

77. Vasilenko SA, Piper ME, Lanza ST, Liu X, Yang J, Li R. Time-Varying Processes Involved in Smoking Lapse in a Randomized Trial of Smoking Cessation Therapies. Nicotine Tob Res. 2014;16: S135-S143. doi:10.1093/ntr/ntt185

78. Cambron C, Lam CY, Cinciripini P, Li L, Wetter DW. Socioeconomic Status, Social Context, and Smoking Lapse During a Quit Attempt: An Ecological Momentary Assessment Study. Ann Behav Med. 2020;54: 141-150. doi:10.1093/abm/kaz034

79. Jahnel T, Ferguson SG, Shiffman S, Thrul J, Schüz B. Momentary smoking context as a mediator of the relationship between SES and smoking. Ambul Assess Addict Disord. 2018;83: 136-141. doi:10.1016/j.addbeh.2017.12.014