Early Attrition Prediction for Web-Based Interpretation Bias Modification to Reduce Anxious Thinking: Machine Learning Study

Sonia Baee, Jeremy W Eberle, Anna N. Baglione, Tyler Spears, Elijah Lewis, Henry C. Behan, Hongning Wang, Daniel H. Funk, Bethany Teachman, Laura E Barnes

Submitted to: JMIR Mental Health
on: August 03, 2023

Table of Contents

Early Attrition Prediction for Web-Based Interpretation Bias Modification to Reduce Anxious Thinking: Machine Learning Study

Sonia Baee[1] BSc, MSc, PhD; Jeremy W Eberle[2] BSc, MSc; Anna N. Baglione[1] BSc, MSc, PhD; Tyler Spears[3] BSc, MSc; Elijah Lewis[4] BSc; Henry C. Behan[2] BA, MSc; Hongning Wang[4] BSc, MSc, PhD; Daniel H. Funk[5] BSc; Bethany Teachman[2] PhD, BSc; Laura E Barnes[1] BSc, MSc, PhD

[1]Department of Systems and Information Engineering University of Virginia Charlottesville US
[2]Department of Psychology University of Virginia Charlottesville US
[3]Department of Electrical and Computer Engineering University of Virginia Charlottesville US
[4]Department of Computer Science and Technology Tsingua University Beijing CN
[5]Sartography Staunton US

Corresponding Author:
Laura E Barnes BSc, MSc, PhD
Department of Systems and Information Engineering
University of Virginia
151 Engineer's Way
Charlottesville
US

Abstract

Background: Digital mental health is a promising paradigm for individualized, patient-driven healthcare. For example, cognitive bias modification programs that target interpretation biases (CBM-I) can provide practice thinking about ambiguous situations in less threatening ways online without requiring a therapist. However, digital mental health interventions, including CBM-I, are often plagued with lack of sustained engagement and high attrition rates. New attrition detection and mitigation strategies are needed to improve these interventions.

Objective: The present analyses aimed to identify participants at high risk of dropout during the early stage of three web-based trials of multi-session CBM-I and to investigate which self-reported and passively detected feature sets from the intervention and assessment data were most informative in making this prediction.

Methods: Participants were community adults with trait anxiety or negative future thinking (Study 1 N = 252, Study 2 N = 326, Study 3 N = 699) who had been assigned to CBM-I conditions in three efficacy-effectiveness trials on our team's public research website. To identify participants at high risk of dropout, we created four unique feature sets: self-reported baseline user characteristics (e.g., demographics), self-reported user context and reactions to the program (e.g., state affect), self-reported user clinical functioning (e.g., mental health symptoms), and passively detected user behavior on the website (e.g., time spent on a web page of CBM-I training exercises; time of day; latency of completing assessments; type of device used). Then, we investigated the feature sets as potential predictors of which participants were at high risk of not starting the second training session of a given program using well-known machine learning algorithms.

Results: The extreme gradient boosting algorithm (XGBoost) performed the best and identified high-risk participants with F1-macro scores of .832 (Study 1 with 146 features), .770 (Study 2 with 87 features), and .917 (Study 3 with 127 features). Features involving passive detection of user behavior contributed the most to the prediction relative to other features (mean Gini importance scores and 95% CIs = .033 ± .014 in Study 1; .029 ± .006 in Study 2; .045 ± .006 in Study 3). However, using all features extracted from a given study led to the best predictive performance.

Conclusions: These results suggest that using passive indicators of user behavior, alongside self-reported measures, can improve prediction of participants at high risk of dropout early in the course of multi-session CBM-I programs. Further, our analyses highlight the challenge of generalizability in digital health intervention studies and the need for more personalized attrition prevention strategies.

(JMIR Preprints 03/08/2023:51567)
DOI: https://doi.org/10.2196/preprints.51567

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Early Attrition Prediction for Web-Based Interpretation Bias Modification to Reduce Anxious Thinking: Machine Learning Study

SONIA BAEE, University of Virginia JEREMY W. EBERLE, University of Virginia ANNA N. BAGLIONE, University of Virginia TYLER SPEARS, University of Virginia ELIJAH LEWIS, University of Virginia HENRY C. BEHAN, University of Virginia HONGNING WANG, Tsinghua University DANIEL H. FUNK, Sartography, Staunton, VA BETHANY A. TEACHMAN, University of Virginia LAURA E. BARNES, University of Virginia

Background: Digital mental health is a promising paradigm for individualized, patient-driven healthcare. For example, cognitive bias modification programs that target interpretation biases (CBM-I) can provide practice thinking about ambiguous situations in less threatening ways online without requiring a therapist. However, digital mental health interventions, including CBM-I, are often plagued with lack of sustained engagement and high attrition rates. New attrition detection and mitigation strategies are needed to improve these interventions. *Objectives:* The present analyses aimed to identify participants at high risk of dropout during the early stage of three web-based trials of multi-session CBM-I and to investigate which self-reported and passively detected feature sets computed from the participant interacting with the intervention and assessments were most informative in making this prediction. *Methods:* Participants were community adults with trait anxiety or negative future thinking (Study 1 N = 252, Study 2 N = 326, Study 3 N = 699) who had been assigned to CBM-I conditions in three efficacy-effectiveness trials on our team's public research website. To identify participants at high risk of dropout, we created four unique feature sets: self-reported baseline user characteristics (e.g., demographics), self-reported user context and reactions to the program (e.g., state affect), self-reported user clinical functioning (e.g., mental health symptoms), and passively detected user behavior on the website (e.g., time spent on a web page of CBM-I training exercises; time of day; latency of completing assessments; type of device used). Then, we investigated the feature sets as potential predictors of which participants were at high risk of not starting the second training session of a given program using well-known machine learning algorithms. *Results:* The extreme gradient boosting algorithm (XGBoost) performed the best and identified high-risk participants with F1-macro scores of .832 (Study 1 with 146 features), .770 (Study 2 with 87 features), and .917 (Study 3 with 127 features). Features involving passive detection of user behavior contributed the most to the prediction relative to other features (mean Gini importance scores and 95% CIs = .033 ± .014 in Study 1; .029 ± .006 in Study 2; .045 ± .006 in Study 3). However, using all features extracted from a given study led to the best predictive performance. *Conclusions:* These results suggest that using passive indicators of user behavior, alongside self-reported measures, can improve prediction of participants at high risk of dropout early during multi-session CBM-I programs. Further, our analyses highlight the challenge of generalizability in digital health intervention studies and the need for more personalized attrition prevention strategies.

INTRODUCTION

Approximately half of the U.S. population will experience a mental illness during their lifetime [93, 95]. During the early stage of the COVID-19 pandemic, researchers estimated an increase of 25.6% in new cases of anxiety disorders per 100,000 people, globally [88]. Mental illness is associated with impaired daily functioning, more frequent use of healthcare resources, and increased risk of suicide [93]. However, more than two thirds of individuals with a mental illness do not receive treatment [67]. A multitude of barriers impede the initiation and sustained use of face-to-face (i.e., traditionally delivered) treatment, including stigma; cost; lack of insurance coverage; and limited availability of support services, especially trained clinicians [6, 71, 72, 95]. Given these challenges, there is an urgent need to help people manage their mental health in new ways [6, 93].

Digital mental health interventions (DMHIs), which harness digital technologies to promote behavior change and maintain health [110], provide an appealing alternative for much-needed treatment outside a clinician's office [78]. DMHIs may help individuals to overcome obstacles to treatment, such as geographic or financial constraints, and may thus reduce the treatment gap across the broader population. Given the limited resources for healthcare service delivery, low-cost mobile health (mHealth) and electronic health (eHealth) interventions could be key to supporting symptom monitoring and self-management of patients with mental disorders over time [10]. With increasing demand for mental health care amid a shortage of mental health professionals, the use of eHealth and mHealth applications is expanding [116, 118, 120]. While these solutions have the potential to play an important role in increasing access to mental health services, especially for underserved communities, the clinical community is still determining how to best leverage these solutions [117].

Poor adherence and substantial dropout in DMHIs are common challenges in DMHIs [115]. *Adherence*, the extent to which users complete a DMHI's tasks as they were intended [63, 75], is likely associated with better treatment outcomes [93], for although these tasks can vary widely (given the varied designs of DMHIs; [63]), it is through engaging with these tasks that DMHIs are thought to achieve their outcomes [75]. However, sustained engagement with these platforms remains a significant issue [10, 33, 49, 58, 94, 97]. *Dropout*, which occurs when a participant prematurely discontinues an intervention (for many potential reasons–e.g., technical issues, lack of time or energy, lack of perceived benefit; [33]), ranges from 30% to 90% in digital health interventions [2, 6, 49, 61, 93, 101]. Even a modest dropout rate can limit the generalizability of digital intervention findings to only those who completed the study; thus, effective evaluation of treatments becomes a challenge [10, 44, 59, 61, 70, 94]. This likely contributes to the uncertainties among clinicians and patients regarding the efficacy, usability, and quality of DMHIs [10]. While there are many reasons clinicians tend not to integrate DMHIs into their clinical practice (e.g., insufficient knowledge about DMHIs and lack of training about how to integrate them; [[123]), if patients' sustained engagement with DMHIs is low and they stop before achieving meaningful gains, then clinicians have little incentive to view them as a helpful tool to increase efficiency and impact of care.

One approach to reducing attrition in DMHIs is to identify participants at high risk of dropping out at the early stages of the DMHI so that the intervention can be adapted to these users' needs [26], or so that more support (e.g., minimal human contact with a telecoach) can be offered specifically to such users (thereby maintaining scalability, [121]). Although increasing attention has recently been dedicated to attrition in various eHealth interventions [22, 62, 106], relatively few advances within DMHIs have predicted dropout through streamlined quantitative approaches considering both passive and self-report data. Testing the effectiveness of interventions on treatment outcomes [48] often takes priority over identifying and predicting users at high risk of attrition. Consequently,

methodological advancements in attrition prediction have largely taken place outside clinically relevant settings, such as in the eCommerce and social gaming industries [19, 54, 83]. The present study develops a data-driven algorithm that includes both passive indicators of user behavior and self-report measures to identify individuals at high risk of early attrition in three DMHIs; as such, it provides a framework that may inform adaptation of DMHIs to be more personalized to an individual user based on attrition risk.

To predict attrition in DMHIs, there are two main considerations [33]. First, we need to define the prediction horizon; that is, researchers should determine the point in an intervention's timeline at which it would be beneficial to predict which participants are at high risk of dropping out. This decision may be informed by an analysis of when in the timeline most participants are actually dropping out; such an analysis may allow the identification and strengthening of weak parts of an intervention. This decision may also be informed by considering typical patterns of engagement, given that low engagement has been consistently cited as the construct underlying attrition [10, 19, 39, 41, 54, 84, 86, 105]. However, engagement is a very broad construct with many components [75], and empirical evidence suggests that engagement fluctuates over time [26]. Thus, carefully defining the feature space and predicting participants at high risk of attrition at meaningful time points in a program can provide valuable information. For example, participants may stay in the intervention first out of curiosity, which relates to the novelty effect—the human tendency to engage with a novel phenomenon [48], but then lose interest. If a researcher wants to mitigate the impact of the novelty effect, for example, then understanding early-stage dropout (i.e., early in the program but once it is no longer brand new and unknown) is critical.

Second, we must consider which factors cause users to drop out of a given DMHI. Answering this question can help researchers and designers tailor the intervention to particular user groups. Demographic variables such as gender, age, income, and educational background have been related to higher attrition rates in digital health interventions [10, 29, 74, 79, 85]. With respect to participants' mental health (e.g., lifetime symptoms assessed at baseline or current symptoms assessed over the course of the intervention), the presence of symptoms may increase interest in and use of a digital intervention in efforts to reduce symptoms [60]. However, certain symptoms (e.g., hopelessness) may reduce motivation or ability to sustain engagement with an intervention [8, 10, 58, 61, 62, 97]. In addition to these baseline user characteristics, user clinical functioning (i.e., current symptoms and psychological processes thought to maintain symptoms), self-reported user context and reactions to interventions (e.g., perceived credibility of DMHIs, which is associated with increased engagement and reduced dropout; [10]), and passively detected user behavior influence attrition rates in digital platforms [61, 92]. This behavior includes time spent using an intervention [18, 53, 83], the passively detected context (e.g., time of day, day of week) [18], and type of technology (e.g., web, smartphone, computer-based, wearable) [58, 107].

Prior studies, mainly in psychology, have predicted attrition primarily with statistical techniques such as ANOVA and regression [34, 45, 80, 85, 91]. In addition, some other research has used macro- level approaches, such as contrasting one intervention's attrition rate against another's [39] and examining participant and psychotherapy trial factors that predict dropout rates [25]. Researchers in computer and data science and the mobile gaming industry have found success in predicting attrition ("churn") using more advanced techniques, which more commonly leverage passively collected behavioral data, such as linear mixed modeling [54], survival analysis [83], and probabilistic latent variable modeling [19]. More recently, advanced machine learning models, such as deep neural networks, have also been useful for modeling and predicting attrition in mobile gaming [53, 83, 96, 111] and in digital health care applications [58, 99]. Our approach builds on work predicting attrition in DMHIs [54, 74, 85, 90, 91, 99, 100] and incorporates both passively

collected behavioral data and self-report data [55, 75, 87, 92, 93, 100, 102].

One attractive DMHI for anxiety is cognitive bias modification for interpretation (CBM-I; [65, 69]), a web-based program with potential to reach large, geographically diverse samples of anxious adults. CBM-I aims to shift threat-focused interpretation biases wherein anxious people tend to assign a negative or catastrophic meaning to situations that are ambiguous. Cognitive models of anxiety suggest that training anxious people to consider benign interpretations of ambiguous situations, as opposed to only rigidly negative interpretations, may reduce anxiety [3, 4, 68]. To shift interpretation biases, CBM-I training sessions prompt users to imagine themselves in ambiguous, threat-relevant scenarios (presented in a set of short sentences) and to practice disambiguating each scenario by filling in its final word (typically presented as a word fragment) [69]. Active CBM-I conditions encourage more positive and flexible interpretation of scenarios by making the final word assign a benign or positive meaning to the ambiguous situation (e.g., "As you are walking down a crowded street, you see your neighbor on the other side. You call out, but she does not answer you. Standing there in the street, you think that this must be because she was distrac_ed."). By presenting benign or positive endings for most scenarios (e.g., 90%), positive CBM-I conditions train a positive contingency in that users learn to expect that ambiguous potentially threatening situations usually work out okay. The most improvement is expected in positive conditions relative to other active conditions (e.g., 50% benign/positive-50% negative conditions that present benign and

Table 1. Overview of MindTrails studies.

Study name	Duration	Target population	# CBM-I training sessions	# Valid participants in parent study	# Positive CBM-I participants [a]	Engagement strategy	
						Compensation	Session reminder
Managing Anxiety (MA)	Jun 8, 2016 - Jan 20, 2019	Anxious adults	8	807	252	$0	Emails
Future Thinking (FT)	May 3, 2017 - Oct 16, 2019	Adults with negative expectations about the future	4	1,221	326	$0	Emails, text
Calm Thinking (CT)	May 18, 2019 - Nov 13, 2020	Anxious adults	5	1,748	699	$25 [b]	Emails, text

Note. CBM-I = cognitive bias modification for interpretation.
[a] Condition of interest for present analyses.
[b] $5 per assessment at baseline, after Session 3, and after Session 5; $10 for follow-up assessment.

negative endings in equal proportions, thereby training flexible interpretation but no contingency) and to control conditions (e.g., no training or a neutral condition with emotionally unambiguous scenarios and neutral endings). Thus, the present analyses focus on attrition in positive conditions. Despite some mixed results [87, 108], a number of studies have shown the effectiveness of positive CBM-I conditions for shifting interpretation biases and reducing anxiety symptoms [29, 36, 43, 49, 51, 65]. To benefit from CBM-I programs, participants must be able to use them effectively over a sustained period. However, like many DMHIs, web-based CBM-I programs face substantial attrition rates [46, 49].

The aims of this paper are threefold. The first aim is to determine a practical attrition prediction horizon (i.e., through which session we need to identify individuals at high risk of dropping out). The second aim is to identify participants at high risk of dropping out by leveraging baseline user characteristics, self-reported user context and reactions to the program, passively detected user behavior, and clinical functioning of users within our analysis. The third aim is to explore which of these feature sets are most important for the identification of high-risk participants. To achieve these

aims, we propose a multi-stage pipeline to identify participants at high risk of dropout from the early stage of three different DMHI studies. These interventions all use internet-delivered CBM-I ([65, 69]) to help individuals change their thinking in response to situations that make them feel anxious or upset ([29, 30, 49]). Note that our proposed pipeline is expected to apply broadly to DMHIs; however, we focus on CBM-I programs in this paper as a useful starting point and look for important features of attrition in such programs.

METHODS

Data Source and Interventions

MindTrails (https://mindtrails.virginia.edu) is a multi-session internet-delivered CBM-I training program. To date, over 6,000 people across more than 80 countries have enrolled in MindTrails, pointing to participant interest in accessing a technology-delivered, highly scalable intervention that can shift anxious thinking in a targeted and efficient way.

In this project, we focus on three MindTrails studies: Managing Anxiety (MA), Future Thinking (FT), and Calm Thinking (CT). We provide a brief overview of these studies, which were approved by the University of Virginia Institutional Review Board. We analyzed data from 1,277 participants across these studies. Details of the studies are provided in Table 1.

Participants and Procedure

Study 1: Managing Anxiety. The Managing Anxiety (MA) study (IRB #2703) developed an infrastructure to assess the feasibility, target engagement, and outcomes of a free, multi-session web-based CBM-I program for anxiety symptoms. A large sample of at least moderately trait anxious community adults based on an anxiety screener (Anxiety Scale of the Depression Anxiety Stress Scales, DASS-21; [64]) was randomly assigned to (a) positive CBM-I training (90% positive-10% negative), (b) 50% positive-50% negative CBM-I training, or (c) a no-training control condition. They also underwent an imagery prime manipulation–an imagination exercise toward the start of CBM-I training designed to activate participants' anxious thinking about a situation in their life. After consenting and enrolling, participants completed a battery of baseline measures, including demographic information, mental health history, and treatment history. For details about the MA study protocol, including the study aims and outcome measures, see [49].

The program involved up to eight online training sessions, delivered at least 48 hours apart with assessments immediately after each session, and a 2-month follow-up assessment. CBM-I training at each session involved 40 training scenarios, which were designed to take about 15 min to complete. Study contact, in the form of automated reminder emails sent to all participants, was equivalent in content and schedule regardless of training condition, and if participants completed only part of an assessment task, they continued it the next time they returned. (If participants completed only part of a training task, they restarted the task upon returning.) Participants received no monetary compensation. A total of 3,960 participants completed the eligibility screener, out of which 807 were eligible, enrolled, and completed the baseline assessment. In the present analyses, only data from the positive intervention arm (i.e., positive CBM-I condition) were used (N = 252) given our interest in testing predictors of attrition in positive CBM-I across all three studies.

Study 2: Future Thinking. The Future Thinking (FT) study (IRB #2690), a hybrid efficacy-effectiveness trial, tested a multi-session, scalable web-based adaptation of CBM-I to encourage healthier, more positive future thinking in community adults with negative expectations about the

future based on the Expectancy Bias Task (shortened from [76]). After completing the screener, eligible participants consented, enrolled, and were randomly assigned to (a) positive conditions with ambiguous future scenarios that ended positively, (b) 50-50 conditions that ended positively or negatively, or (c) a control condition with neutral scenarios. For details about the FT study aims and outcome measures, see [29].

Participants were asked to complete four training sessions (40 scenarios each). Assessments were given at baseline, immediately after each session, and at 1-month follow-up. Participants had to wait 2 days before starting the next training session and 30 days before starting the follow-up assessment. Participants had the option of receiving an email or text reminder when the next session or follow-up assessment was due, and if they completed only part of a training or assessment task, they continued it the next time they returned. Participants received no monetary compensation. A total of 4,751 participants completed the eligibility screener, out of which 1,221 were eligible and enrolled. In the present analyses, only data from the positive CBM-I intervention arm (i.e., Positive and Positive + Negation conditions) were used (N = 326).

Study 3: Calm Thinking. The Calm Thinking (CT) study (IRB #2220), a sequential, multiple-assignment, randomized trial (SMART), tested the effectiveness of positive CBM-I relative to a psychoeducation comparison condition (randomly assigned at Stage 1). It also tested the addition of minimal human contact (i.e., supplemental telecoaching randomly assigned at Stage 2, [103]) for CBM-I participants at higher risk of dropout in early stages. Additional details can be found in the main outcomes paper [30].

After completing the anxiety screener (DASS-21-AS), eligible participants consented and enrolled. Participants were asked to complete a baseline assessment; one training session per week over 5 weeks (five sessions, 40 scenarios each in CBM-I), with an assessment immediately after each session; and a 2-month follow-up assessment. If participants completed only part of a training or assessment task, they continued it the next time they returned. Participants were compensated via e-gift cards (see Table 1 for details). A total of 5,267 participants completed the eligibility screener, out of which 1,748 were eligible and enrolled. Cleaned data [28] from the CBM-I-only intervention arm (N = 699; i.e., CBM-I condition excluding high-risk participants randomized to receive supplemental coaching) were used in the present analyses, to allow a clean analysis of attrition during positive CBM-I.

In total, 252 MA participants, 326 FT participants, and 699 CT participants were in the positive CBM-I intervention arm of these studies.

Definition of Attrition. In the present analyses, we predict attrition in multi-session DMHIs. Eysenbach defined two types of attrition [33]: (a) *nonusage attrition*, which refers to participants who stopped using the intervention (i.e., not completing the training sessions), and (b) *dropout attrition*, which refers to participants who were lost to follow-up because they stopped completing research assessments (e.g., did not complete follow-up assessment). In MindTrails studies, training and assessment tasks are intermixed and must be completed in series (e.g., participants cannot complete Session 1 assessment until they complete Session 1 training, they cannot complete Session 2 training until they complete Session 1 assessment, and so on). Due to this sequential design, nonusage and dropout attrition are conflated in our studies. Since it is impossible to skip any training or assessment tasks, in this paper we simply use the term *attrition*.

Ethical Considerations

All three studies were reviewed and approved by the Institutional Review Board (IRB) of the University of Virginia. After screening, eligible participants provided informed consent for “a new

internet-based program designed to reduce anxiety." Data were stored in accordance with University of Virginia Information Security policies, and de-identified data were analyzed. Participants were compensated up to $25 in e-gift cards: $5 for each assessment at pretreatment and after Sessions 3 and 5, and $10 for the follow-up assessment.

ATTRITION PREDICTION

Pipeline

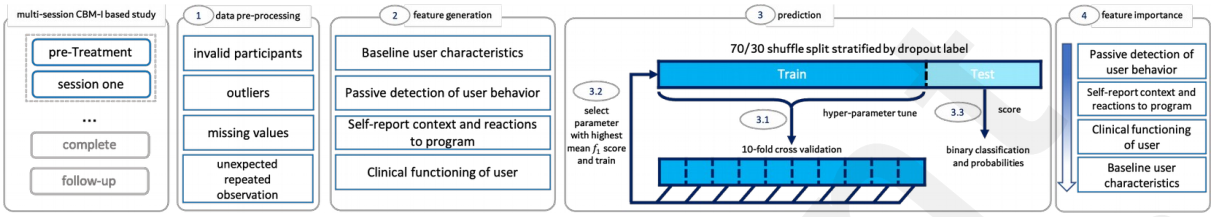


Fig. 1. Overview of pipeline predicting early-stage attrition in web-based multi-session cognitive bias modification for interpretation (CBM-I) interventions.

DMHIs are often divided into multiple phases, sometimes called *modules*. In this paper, we refer to modules as *sessions* to mirror the language used by mental health specialists for in-person treatment (e.g., holding “sessions” with a client). We proposed a pipeline that is built to handle multi-session DMHI datasets with a diverse set of features. We also assumed the study contained one or more assessment or training sessions to achieve the study goals since our focus is on multi-session studies. Therefore, we required at least one observation from each participant for the selected features.

Predicting early-stage dropout in DMHIs is challenging and requires several key tasks. We first determined the prediction horizon of the selected CBM-I interventions (Aim 1). We then organized the remaining tasks into four main steps from the data science and engineering literature: (1) data preprocessing, (2) feature generation, (3) predictive modeling, and (4) feature importance. We outline these steps in the context of attrition prediction in DMHI in Figure 1 and describe each step below.

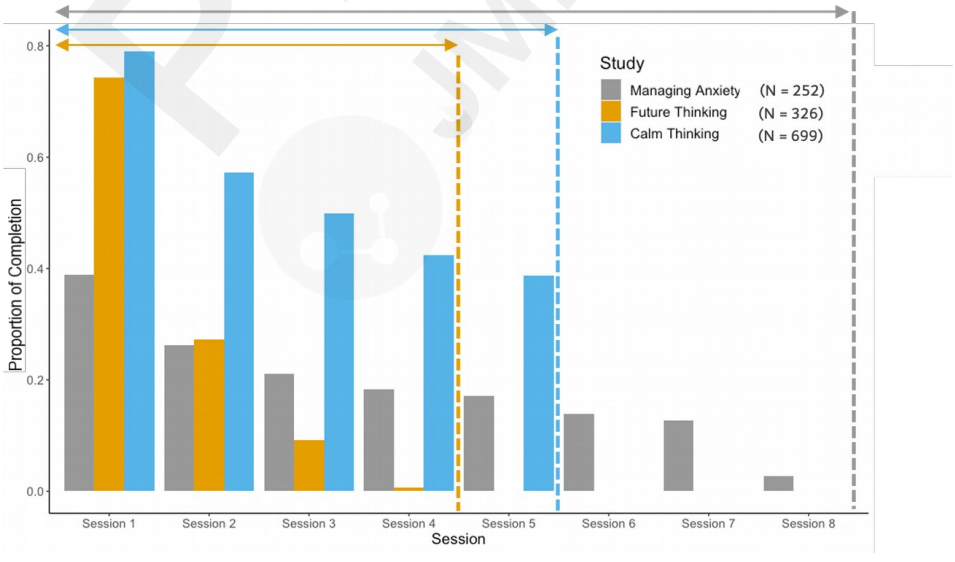


Fig. 2. Proportion of completion per training session (out of participants who started Session 1 training) by study. The session was deemed completed if participants completed the last questionnaire in the assessment that immediately followed the training session. Dashed lines show

the last training session for each study.

Prediction Horizon

To analyze when users stopped using the intervention (Aim 1), the proportions of participants who completed each training session (out of the number of participants who started Session 1 training) were visualized (see Figure 2). In this figure, each session was considered completed if participants completed (i.e., had an entry in the Task Log for) the last questionnaire in the assessment that immediately followed a given training session. We decided to predict which participants who had started Session 1 training were at high risk of dropping out before starting Session 2 training (Aim 2) for the following reasons. First, our goal is to make inferences about user dropout during DMHIs (not simply using baseline assessments to predict who does not even start the program). We restricted the sample to participants who had started Session 1 training because we consider these participants as part of the intent-to-treat (ITT) sample. Second, the highest rate of attrition was observed between starting the first training session and finishing the second session's assessment, with most dropout occurring between sessions (vs. during Session 1 or Session 2). For this reason, we wanted to predict participants at high risk of dropping out before starting Session 2 training. Notably, identification of participants who are at high risk of dropping out early in the program might decrease the attrition rate at the end of the intervention by detecting high-risk participants sooner rather than later so targeted supports can be added to increase retention at pivotal times.

Data Preprocessing

All data must be *preprocessed* before analysis, especially data collected outside a controlled lab environment. In the following paragraphs, we describe our methods for addressing issues such as invalid participant data, outliers, and missingness during preprocessing.

Invalid Participants. One of the main challenges in online digital mental health studies is to distinguish spam and bot-generated responses from real responses [28, 49]. Malicious actors often employ bots to complete questionnaires when they learn of an appealing incentive, such as monetary compensation for participating in a study. To increase the validity of the input data, we removed any suspicious responses, such as those that were submitted quickly (e.g., less than 5 seconds for half of all questions in a given measure) or contained submissions that violated the required wait time (e.g., 48 hours) between sessions.

Outliers. To reduce the likelihood of identifying coincidental events, we first normalized the data by using the Z-score metric. We then identified and removed outliers; since we did not expect to have very large or small data values [9], we excluded outliers at least three standard deviations from the mean value [73] for numerical variables and used visual inspection of a frequency distribution, a histogram with Freedman-Diaconis rule to determine the bin width, for categorical variables.

Missing Values. Real-world data collection is often messy; technical issues, dropout, and loss of network connection are all common issues that arise and can lead to missing values for some or all items of a given questionnaire. In addition, participants in DMHIs are often given the option to decline to answer items when responding to a self-report questionnaire. This may be done either implicitly, in that the question is not required, or explicitly, in that the participant is given a set of options where one is "prefer not to answer" (or similar response). The challenge with empty or "prefer not to answer" values is that they both function as missing values.

Missing values are a fundamental issue in digital health interventions for several reasons [113]. Most machine learning techniques are not well prepared to deal with missing data and require that the data be modified through imputation or deletion of the missing records. Additionally, missing data may significantly impact the predictive analysis, as well as descriptive and inferential statistics [40]. To address these issues, we employed several imputation approaches to handle the challenge of missing data in some or all items in the required features and time points for different types of variables. Without imputation, this missing data could lead to more bias, decreased statistical power, and lack of generalizability [113].

We handled missing data for all features, for each unique time point, using the following methods: Out of the initial set of features (221 for MA, 109 for FT, 241 for CT), we first removed features/variables at a given time point that have missing values in more than 80% of all valid participants. The percentages of features removed for this reason in MA, FT, and CT studies were 33.94%, 20.18%, and 47.30%, respectively, yielding a final set of 146, 87, and 127 features, respectively. Next, we imputed categorical variables at a given time point with the most frequent answers at that time point of participants with the same demographics. To do so, we grouped participants based on two of the demographic characteristics (i.e., education and gender, which were the most complete). To impute the numerical individual item variables at a given time point, we utilized the *k*-nearest neighbors (KNN) method [66] to replace the missing values in the same demographic group with the mean value at that time point from the five nearest neighbors found in the training set. We used a Euclidean distance metric [52] to impute the missing values.

Unexpected Multiple Observations. Unexpected multiple observations may be present within a DMHI dataset for several reasons. Participants might complete the eligibility screener multiple times to gain access to the intervention if they were previously screened out or to achieve a more desirable score. Technical issues can also cause duplicate values. For example, a brief server error may cause a questionnaire to be submitted more than once. We used one of the following two strategies to handle unexpected multiple observations: (a) calculate the average values of each item across the observations, or (b) keep the latest observation. We selected one of the strategies based on the temporal latency between unexpected multiple observations. If the temporal latency between unexpected multiple observations was less than the mean latency across all participants, we applied the first strategy. Otherwise, the second strategy was selected.

Feature Generation

Baseline User Characteristics. (Note: Measures without citations in this section and the sections below were developed by the MindTrails research team.) At the baseline assessment of the three studies, the following demographic variables were assessed: age, gender, race, ethnicity, education, employment status, marital status, income, and country. History of mental health disorders and treatment were also assessed. The MA and CT studies also asked participants about the situations that make them anxious, called *anxiety triggers*. We included these measures in our baseline user characteristics feature set (see details in Table 2).

Self-Reported User Context and Reactions to Program. The importance of reducing anxiety or changing thinking (Importance Ruler, modified from [102]) and confidence in the intervention (modified from [11]) were assessed at the baseline assessment of every study. In addition, after completing a given session's assessment, participants were asked for the date they would return for the next session. State anxiety (in MA and CT; Subjective Units of Distress, SUDS, modified from [104]) or current positive and negative feelings (in FT) were assessed before and after participants completed each session's training. The MA and CT studies also assessed participants' peak anxiety when imagining an anxiety-provoking situation in their lives as part of the anxious imagery prime

completed toward the start of training. At the end of each session in the CT study, the participant's location, level of distraction, and ease of use of the program were assessed. All of these measures were included in the self-reported user context and reactions to the program feature set (see details in Table 2).

Passive Detection of User Behavior. To further understand participants' context and behavior when interacting with the platform, the following variables were calculated: time spent on a page, time of day, day of the week, and latency of completing assessments. Type of device (i.e., desktop, tablet, smartphone) was also included as a feature given that multiple devices could be used to access the program, each with different characteristics (e.g., screen size, input methods, mobility) that could influence user behavior. In most cases, these variables were extracted for each assessment and training task for each session. For details about which features were extracted for which studies, see Table 2.

User Clinical Functioning. Primary and secondary outcome measures used to evaluate the effectiveness of the intervention were included in the clinical functioning feature set. These measures assessed interpretation bias (Recognition Ratings, RR, modified from [69]; Brief Body Sensations Interpretation Questionnaire, BBSIQ, modified from [20]), expectancy bias (Expectancy Bias Task, modified from [76]), anxiety symptoms (Overall Anxiety Severity and Impairment Scale, OASIS, adapted from [82]; DASS-21-AS; Generalized Anxiety Disorder-2 scale, GAD-2, modified from [56]), and comorbid depression symptoms (DASS-21-DS; Patient Health Questionnaire-2, PHQ-2, modified from [55]) and alcohol use (Daily Drinking Questionnaire, DDQ, [21]). They also assessed the centrality of anxiety to identity (Anxiety and Identity Circles, modified from [32]) and other cognitive mechanisms, including cognitive flexibility (Cognitive Flexibility Inventory, CFI, adapted from [24]), experiential avoidance (Comprehensive Assessment of ACT Processes, CompACT, modified from [37]), cognitive reappraisal (Emotion Regulation Questionnaire, ERQ, modified from [42]), and intolerance of uncertainty (Intolerance of Uncertainty Scale-Short Form, IUS-12, modified from [12]). Finally, they assessed self-efficacy (New General Self-Efficacy Scale, NGSES, modified from [14]), growth mindset (Personal Beliefs Survey, PBS, modified from [27]), optimism (Life Orientation Test-Revised, LOT-R, modified from [89]), and life satisfaction ([17]; Quality of Life Scale, QOL, [35]). For details about which features were extracted for which studies, see Table 2.

Predictive Modeling

For each study, predictors of attrition were investigated after participants started Session 1 training, imputing any missing values for features collected during Session 1 training or assessment.

To identify participants at high risk of dropping out before starting the second training session, the following predictors of attrition were investigated: baseline user characteristics (at the pretest assessment), self-reported user context and reactions to the program, passively detected user behavior, and clinical functioning of users. We used data from the pretest, the first training session, and the assessment following the first training session.

Dropout Label. For each participant, we calculated a binary ground truth label for their actual dropout status before starting the second training session, where 0 indicates the participant started training for the second session and 1 indicates the participant did not start training for the second session (i.e., dropped out). A participant was deemed as having started a given session's training if they had an entry in the Task Log for the Affect task, which was administered immediately before the first page of training materials for each session.

Class Imbalance. Class imbalance is a common problem for supervised learning tasks such as attrition prediction. Such datasets have one or more classes (e.g., “did not dropout” in the case of CT) that have a greater number of observations than other classes (e.g., “dropped out” in CT). Class imbalance can worsen the performance of machine learning models by biasing them toward learning the more commonly occurring classes. We used the synthetic minority oversampling technique (SMOTE) [13] to help rectify the class imbalance.

SMOTE resolves this challenge by generating synthetic samples for the minority class, with the aim of balancing the distribution of samples between the two classes. The technique operates by selecting two or more samples from the minority class and computing the difference between their features. This difference is then added to the feature values of one of the selected samples to create a new synthetic sample. This process is repeated to generate a sufficient number of synthetic samples, which are then added to the original dataset to achieve an optimal balance between the majority and minority classes. It has proven to be very effective in dealing with class imbalance problems for tabular datasets [31] (see Figure 2).

Classification. Binary classification is a well-studied problem in the machine learning literature [5, 57], and a plethora of models and approaches exist for predicting attrition. We selected leading machine learning models, beginning with simpler, more interpretable models, and progressing to more expressive models for identifying the best predictors of early-stage dropout. Choosing this range of models ensures that our results are not an artifact of an a priori selection of specific model types. We trained and validated a range of models, described in detail below and listed in Table 3. Models that learn a linear decision boundary are typically the first approach for binary classification problems. These models separate participants into two classes defined by the estimated decision boundary, in our case participants who drop out and those who remain. The logistic regression model estimates this decision boundary by minimizing the mean squared error of predictions in the training set [47]. Similarly, the support vector machine (SVM) estimates this boundary by maximizing the distance from the “edge” of each class. Some non-linearity is also introduced into the SVM by projecting its feature space with the radial basis function (RBF) kernel [81].

Other models estimate a non-linear decision boundary. A decision tree model estimates a continuous piecewise boundary, with each “piece” indicating a different set of conditions that leads to a particular leaf node of the tree [15]. We further evaluated several tree-based ensemble models. In ensemble models, multiple sub-models are composed to form a prediction. The random forest model uses decision trees as its sub-model, creating a “forest” (set) of such trees. The random forest estimates the best feature subset to give to each tree while maximizing the average prediction accuracy over all trees [15]. Similarly, AdaBoost comprises multiple shallow decision trees, giving a weighting to each tree according to the overall prediction accuracy [15].

Finally, gradient boosting algorithms (and the related extreme gradient boosting method, XG- Boost [16]) were used to train ensembles of decision trees. Gradient boosting minimizes an objective function that is differentiable with respect to all sub-model parameters, and the sub-model parameters are adjusted via gradient descent. XGBoost [16] is based on the same concept, but also includes parameter regularization to prevent overfitting, and second-order derivatives to control gradient descent. The regularized greedy forest (RGF) model was also evaluated. RGF not only includes tree-structured regularization learning, but also employs a fully corrective regularized greedy algorithm [50]. Finally, a multi-layer perceptron (MLP) model was used. This neural network model implements a feed-forward architecture that backpropagates error with stochastic gradient descent [38].

We employed 10-fold cross-validation stratified by dropout label (i.e., dropout vs. not-dropout) across 100 iterations. Hyperparameter tuning was performed using group five-fold cross-validation on the training set. Hyperopt [7] was utilized to optimized hyperparameters including number of estimators, learning rate, maximum tree depths, C parameter, and gamma. We evaluated up to 30 combinations of these parameters to maximize the model's average macro-F1 score across five folds. The set of hyperparameters that achieved the highest average macro-F1 score across the five folds was chosen to train the model on the entire training set during the outer split.

Model Optimization. To enhance model performance and efficiency, optimization techniques were applied. For instance, in the SVM model, we selected the RBF kernel with gamma determined as 1/(number of features × $X.var()$) to control the influence of training examples. In decision tree models, all features were considered for finding the best splits, while feature subsampling was employed to reduce model correlation and variance.

Our selected criterion for the decision model is entropy, which measures the degree of disorder of the features in relation to the target. The optimum split is chosen by the feature with the lowest entropy. It gets its maximum value when the probability of the two classes is the same. A node is pure when the entropy has its minimum value, which is 0. For the random forest model, we take all the features that make sense in every tree.

In the XGBoost model, we set the subsample ratio of columns for each level equal to 0.4. The subsampling occurs once for every new tree. The γ parameter in XGBoost is used as a threshold for creating new splits in the tree; it represents the minimum loss reduction required to make a further partition on a leaf node of the tree. We set $\gamma = 8$. To control the balance of positive and negative weights in a binary classification problem, we set the parameter $scale\_pos\_weight = sum(\text{negative instances})/sum(\text{positive instances})$. This parameter allows adjustment of the relative weight of positive instances in the cost function, by setting it to the ratio of negative to positive instances. This can help to handle imbalanced datasets where one class is under-represented, as in our case. The eta parameter, learning rate, controls the step size shrinkage used in updating the weights to prevent overfitting. We tuned eta for our models and dataset and got the value 0.01. After each boosting step in XGBoost, we can directly get the weights of newly added features, and eta shrinks the feature weights and the weights of all the features in the model to make the boosting process more conservative. The $\alpha = 0.3$ parameter in XGBoost is used as a regularization term on the weights; it represents the L1 regularization term, which is used to add a penalty term to the cost function that is proportional to the absolute value of the weights. This helps to prevent overfitting by shrinking the weights toward zero. The $\lambda = 0.4$ parameter in XGBoost is also used as a regularization term on the weights; it represents the L2 regularization term, which is used to add a penalty term to the cost function that is proportional to the square of the weights. This helps to prevent overfitting by shrinking the weights toward zero.

For RGF, we used the min-penalty regularization with sum-to-zero sibling constraints to improve the interpretability of the model. For logistic regression, we set the regularization to *elasticnet* and the regularization strength to 1, $C = 1$. For a multi-layer perceptron (MLP), the activation function is set to the rectified linear unit (ReLU) function, represented as $f(x) = max(0, x)$. The initial learning rate for the Adam algorithm is also set to 0.001. It is worth noting that we kept the other hyperparameters of the models at their default values to avoid overfitting and to ensure the stability of the models.

Evaluation Metrics. We used three standard metrics to evaluate attrition prediction: F1-macro, area under the Receiver Operating Characteristic (ROC) curve (AUC), and accuracy. For F1-macro, an F1 score is first computed for each class. The F1 score is the harmonic mean of *precision* (proportion of positive predictions that are correct) and *recall* (proportion of positive classes that are

correctly predicted; *true positive rate*), and it rewards true positives and penalizes false positives and false negatives. F1 scores range from 0 (when no positive predictions are correct) to 1 (when all positive predictions are correct, and no incorrect negative predictions are made). F1-macro is the arithmetic mean of F1 scores across classes and is widely used when classes are imbalanced because it avoids bias towards the majority class by weighting each class's F1 score equally.

AUC, a widely adopted performance metric, measures the trade-off between the true positive rate and the *false positive rate* (proportion of negative classes that are incorrectly predicted as positive) by plotting these rates against one another for various classification thresholds (i.e., probabilities above which a positive prediction is made) and quantifying the area under the resulting ROC curve; this area provides an aggregate measure of performance across all possible thresholds. AUC ranges from 0 (when no positive classes are correctly predicted and all negative classes are incorrectly predicted) to 1 (when all positive classes are correctly predicted and no negative classes are incorrectly predicted), indicating the model's ability to differentiate between positive and negative classes (a value of .5 reflects random prediction).

Accuracy, in turn, is the proportion of all predictions (positive and negative) that are correct and ranges from 0 (no predictions are correct) to 1 (all predictions are correct), providing a straightforward assessment of the model's overall performance, although it can be misleading in isolation when classes are imbalanced. For F1-macro and AUC, scores above .5 are generally considered to reflect reasonable performance, while for accuracy, a score above .7 is considered reasonable.

Feature Importance

Aim 3 of this paper is to explore which feature sets are most important for the identification of high-risk participants. To analyze this, the effect of each feature set on the prediction models was calculated (i.e., Gini importance [77]). We report the mean Gini importance score across two iterations. Gini importance scores reflect the importance of a feature set relative to others (not absolute importance) and can range from 0 to 1, with higher scores reflecting greater importance.

Table 2. Selected features by set extracted from cognitive bias modification for interpretation studies.

Set	Task (From Task Log)	Description	Study	Session
Baseline user characteristics	Demographics	Assesses age, gender, race, ethnicity, education, employment status, marital status, income, country	MA, FT, CT	Baseline
	Mental Health History	Assesses mental health disorders and treatments	MA, FT, CT	Baseline
	Anxiety Triggers	Assesses situations that prompt anxiety	MA, CT	Baseline
Self-reported context and reactions to program	Credibility	Assesses importance of reducing anxiety or changing thinking (Importance Ruler) and confidence in intervention [11]	MA, FT, CT	Baseline
	Return Intention	Assesses days until returning	MA, FT, CT	Session 1
	Affect	Assesses state anxiety (SUDS; in MA and CT) or current positive and negative feelings (in FT)	MA, FT, CT	Session 1
	Impact of Anxious Imagery Prime	Assesses peak anxiety during imagery prime	MA, CT	Session 1
	Session Review	Assesses location, level of distraction, and ease of use of program	CT	Session 1
Passive detection of user behavior	All assessment and training tasks	Computed time on page, time of day, day of week	MA, FT, CT	Baseline, Session 1
	All assessment and training tasks	Computed cumulative time elapsed to complete all components of a given task and latency between completing one task and starting the next	CT	Baseline, Session 1
	Training task (for FT), all assessment and training tasks (for CT)	Device (from Training table for FT, from Task Log for CT)	FT, CT	Baseline, Session 1
	Interpretation Bias (RR)	Assesses positive and negative interpretations of ambiguous situations (each valence scored separately, including both threat-related and threat-unrelated items [a])	MA, CT	Baseline
	Interpretation Bias (BBSIQ)	Assesses positive and negative interpretations of ambiguous situations (each valence scored separately, including items for both internal and external events and excluding neutral items)	MA, CT	Baseline
	Expectancy Bias	Assesses positive and negative expectations for ambiguous future situations (Expectancy Bias Task; each valence scored separately)	FT	Baseline, Session 1
	Anxiety (OA)	Assesses anxiety symptoms (OASIS)	MA, CT	Baseline, Session 1
	Anxiety (DASS21-AS)	Assesses anxiety symptoms (DASS-21-AS)	MA, CT	Baseline
	Anxiety and Depression (PHQ-4)	Assesses anxiety (GAD-2) and depression (PHQ-2) symptoms (each measure scored separately)	FT	Baseline
	Depression (DASS21-DS)	Assesses depression symptoms (DASS-21-DS)	MA	Baseline
	Daily Drinking	Assesses alcohol use (DDQ)	MA	Baseline
	Anxiety Identity	Assesses centrality of anxiety to identity (Anxiety and Identity Circles)	CT	Baseline
	Mechanisms	Assesses cognitive flexibility (CFI), experiential avoidance (CompACT), cognitive reappraisal (ERQ), and intolerance of uncertainty (IUS-12; each measure scored separately)	CT	Baseline
	Wellness (What I Believe)	Assesses self-efficacy (NGSES), growth mindset (PBS), and optimism (LOT-R; each measure scored separately)	FT	Baseline
	Wellness	Assesses self-efficacy (NGSES), growth mindset (PBS), optimism (LOT-R), and life satisfaction ([17]; each measure scored separately)	CT	Baseline
	Wellness (Qual. of Life)	Assesses life satisfaction (QOL)	MA	Baseline

Note. Task Log is a log table that tracks completion of each assessment and training task for each participant in a given study; when the task's content is not evident in the task's name, the content is listed and the name is in parentheses. MA = Managing Anxiety; FT = Future Thinking; CT = Calm Thinking. Other acronyms are defined in section Feature Generation.

[a] Positive and negative interpretation bias assessed using Recognition Ratings (RR) are typically scored using only the

threat-related items, but given that these are only two features, we do not expect this to markedly impact the algorithm.

RESULTS

Model Performance

The results demonstrate that with these predictors (number of features for MA, FT, and CT studies: 146, 86, 127), we were able to identify participants with a high risk of dropping out before starting the second training session of each study (F1-macro score for XGBoost in MA, FT, and CT studies: .832, .770, .917; Table 3). These results show the effectiveness of different feature sets in predicting attrition in the early stages of the DMHIs. Moreover, these results show the superiority of the XGBoost and the random forest models in predicting attrition (see Table 3). XGBoost always places more importance on functional space when reducing the cost of a model, while random forest tries to place more importance on hyperparameters to optimize the model.

Sensitivity to Imputation

To assess the impact of imputation on our prediction models, we conducted an ablation experiment (i.e., systematic removal of a component of the model to test its effect) that eliminated the imputation step from our pipeline. We utilized the XGBoost classification model in this experiment as it demonstrated superior performance throughout our analyses. The results, presented in Table 4, reveal a substantial decrease in performance when imputation using KNN is removed from the pipeline, highlighting the importance of imputation in our prediction models.

Feature Importance

To investigate how the different feature sets affect the performance of attrition prediction, we calculated the average importance score (i.e., weight) for the important features from the selected high-performing classifier after 100 iterations. Overall, a few trends emerged in identifying individuals at high risk of dropout: The passively detected user behavior feature set, and then the self-reported user context and reaction to the program feature set, are consistently more important than the user baseline characteristics and user clinical functioning feature sets for predicting early-stage attrition in a multi-session CBM-I intervention (see Figure 3). More specifically, we found that features involving passive detection of user behavior, such as time spent on a web page of CBM-I training exercises, time of day, latency in completing assessments, and the type of device used, were the most informative predictors of attrition, with mean Gini importance scores (with 95% CIs) across two iterations of .033 ± .014, .029 ± .006, and .045 ± .006 for the MA, FT, and CT studies, respectively (see Figure 3).

Table 3. Performance of attrition prediction models within a given study based on f_1, area under curve, and accuracy scores. The models trained on the Managing Anxiety (MA, [49]), Future Thinking (FT, [29]), and Calm Thinking (CT, [28, 30]) studies and were tested on their respective test sets. Each study-model pairing was evaluated using 10-fold cross-validation, with the best model used for testing.

Data		Evaluation Metric		
	Model	F1-macro↑	Area Under Curve↑	Accuracy↑
Training data: MA		Test data: MA with 146 features		
	Logistic regression	0.698	0.774	0.717
	Support vector machine	0.723	0.802	0.760
	Decision tree	0.555	0.610	0.644
	Random forest	0.819	0.827	0.843
	Gradient boosting	0.802	0.808	0.808
	Extreme gradient boosting	**0.832**	0.848	**0.858**
	Regularized greedy forest	0.794	**0.853**	0.823
	Multi-layer perceptron	0.690	0.772	0.723
Training data: FT		Test data: FT with 87 features		
	Logistic regression	0.682	0.752	0.689
	Support vector machine	0.719	0.787	0.728
	Decision tree	0.688	0.745	0.693
	Random forest	0.767	0.840	0.768
	Gradient boosting	0.758	0.823	0.759
	Extreme gradient boosting	**0.770**	**0.844**	**0.771**
	Regularized greedy forest	0.728	0.817	0.735
	Multi-layer perceptron	0.694	0.778	0.703
Training data: CT		Test data: CT with 127 features		
	Logistic regression	0.878	0.874	0.878
	Support vector machine	0.869	0.861	0.869
	Decision tree	0.786	0.895	0.788
	Random forest	0.914	0.917	0.910
	Gradient boosting	0.901	0.908	0.901
	Extreme gradient boosting	**0.917**	**0.926**	**0.918**
	Regularized greedy forest	0.902	0.908	**0.918**
	Multi-layer perceptron	0.878	0.879	0.878

Note. ↑ indicates that higher values are more desirable for a given metric (which each can range from 0 to 1, with f_1 and area under curve values above .5 and accuracy values above .7 generally considered reasonable; see section Evaluation Metrics for details). The highest values for each metric are in boldface.

Table 4. Sensitivity of attrition prediction model performance to imputation. Ablated versions of the proposed pipeline without imputing missing values are compared to the full pipeline in terms of f_1, area under curve, and accuracy scores. All models used extreme gradient boosting (XGBoost) and trained and were tested on all feature sets of the Managing Anxiety (MA), Future Thinking (FT), and Calm Thinking (CT) studies.

Data	XGBoost Model Version	Evaluation Metric		
		F1-macro↑	Area Under Curve↑	Accuracy↑
MA		146 features		
	No Imputation	0.715	0.801	0.716
	Imputation	**0.832**	**0.848**	**0.858**
FT		87 features		
	No Imputation	0.726	0.796	0.729
	Imputation	**0.770**	**0.844**	**0.771**
CT		127 features		
	No Imputation	0.905	0.904	0.910
	Imputation	**0.917**	**0.926**	**0.918**

Note. ↑ indicates that higher values are more desirable for a given metric (which each can range from 0 to 1, with f_1 and area under curve values above .5 and accuracy values above .7 generally considered reasonable; see section Evaluation Metrics for details). The highest values for each metric are in boldface.

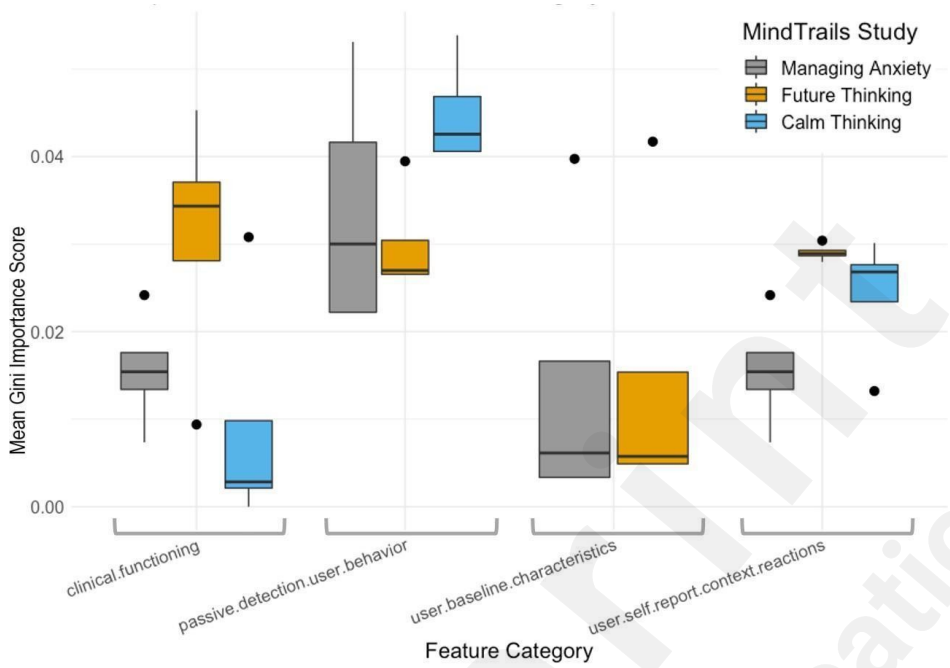


Fig. 3. Importance level of each feature set relative to other feature sets for early attrition prediction in CBM-I studies. Gini importance scores averaged across two iterations are shown. We use XGBoost classifier since it performed the best. These scores reflect the importance of a feature set relative to others (not absolute importance) and can range from 0 to 1, with higher scores reflecting greater importance. Horizontal bars reflect the median score; dots represent outliers, which are observations that fall outside of the boxplot; and whiskers represent the minimum and maximum observations within 1.5 times the interquartile range from the lower and upper quartiles, respectively. Note: No important baseline user characteristic features emerged for the Calm Thinking study.

DISCUSSION

In this research, we investigated the potential of predicting early attrition from three studies of multi-session web-based positive CBM-I training programs by using a combination of features derived from training and assessment data, including baseline user characteristics, self-reported user context and reactions to the program, passive detection of user behavior, and user clinical functioning. Our proposed pipeline was able to identify participants who were at high risk of dropping out early in these studies and provides a framework (i.e., data preprocessing, feature generation, predictive modeling, feature importance) for predicting attrition in DMHIs broadly, although the particulars (e.g., features) will vary with each application. Our results also show that passive features describing user behavior when interacting with a DMHI can be a valuable feature for identifying individuals at high risk of dropping out. In our analyses, interestingly, passive features of user behavior were more informative to this prediction than other features, including user clinical functioning, emphasizing the utility of considering users' real-time behavior in predicting early attrition.

While these findings need to be validated in future studies, they highlight the value of considering collection and use of such features in algorithms for predicting attrition in future DMHI designs. Key next steps include the need to make these data-driven approaches transferable to real-world (non-research) care settings. Clinicians tend not to integrate DMHIs into their clinical practice, in part due to lack of training and understanding about how DMHIs work, which to choose, and how to integrate them [125]. Helping clinicians determine which of their patients is likely to stick with a DMHI (and benefit from it) may help address some of these clinician concerns, and further personalization of the approaches may be useful. Along these lines, more longitudinal features capturing user interaction with DMHIs could enable a level of personalization and customization that has historically been challenging to achieve with only baseline characteristics. It will also be important to address the challenges raised by the complexities of interpreting these algorithms (i.e., determining which factors were key to predicting attrition). When the algorithms seem impenetrable, it may increase clinicians' discomfort with applying them in their practice.

The findings also highlight the value of using *both* passive user behavioral data collected during the DMHI and users' self-report data. Predicting clinical outcomes from single indicators has routinely not been successful. Speaking to the historical challenges in predicting response to depression treatments, van Bronswijk and colleagues noted "no single moderator is likely to be robust enough, on its own, to reliably guide treatment selection…, and indeed none have been identified" [126]. This has led many researchers to recognize the value of novel methods like machine learning that allow for multivariate prediction. The current work extends this approach further by integrating multiple sources of information, beyond only self-report features. This has several advantages, including reducing user burden by not relying solely on self-reported measures, and it allows for prediction to be based on meaningful data about users that they may not have introspective access--or comfort--to report effectively.

Model Performance and Feature Importance

Features extracted from the early stages of a given study (i.e., baseline assessment and Session 1 training/assessment; Table 2) were highly predictive of attrition prior to starting Session 2 training (Table 3). Particularly important was the feature set involving passive detection of user behavior (Figure 3), which consisted of time spent on page, time of day, day of week, time spent on tasks, latency between tasks, and device type. Although it is unclear which passive features were most

informative (a useful future direction), it may be that certain passive features (e.g., time on page) contain real-time information about engagement, motivation, or ability to use the program not captured by other measures (e.g., self-reports of the importance of reducing anxiety or confidence in the program at baseline or self-reports of ease of using the program at end of Session 1). However, the feature importance level varied by classifier and study, highlighting the complexity of identifying individual predictors of attrition. Nevertheless, future studies may benefit from including similar feature sets, especially behavioral features.

Further, our analyses revealed that predicting attrition in DMHI studies is not an easy problem; otherwise, simpler models such as the logistic regression and SVM models may have provided sufficient predictive power. The more complex models that leverage ensembles (random forest, gradient boosting, XGBoost, etc.) performed substantially better without overfitting to the data by making use of cross-validation and parameter tuning. These models are also inherently interpretable, making it easier to explain results to various audiences, including clinicians and other stakeholders. Overall, these results suggest that ensemble/forest models may provide a strong baseline when predicting attrition in CBM-I studies.

Transfer of Knowledge

Given the sparsity of the original dataset, we expected that models would perform better when given informative priors from similar studies. For example, we can use data from the MA study to provide informative priors to the prediction model that is then trained to predict attrition in the CT study. We found that, despite their common use of the MindTrails web infrastructure and use of CBM-I interventions, the three studies (MA, FT, and CT) had substantially different data distributions (i.e., attrition rate and raw values for given features). The studies also had different model performance, not only when each study used all of its own features (Table 3), but also when the studies used only the features they shared (Table 5). Thus, although our findings provide insights into next steps for this research, their generalizability to other CBM-I studies and DMHIs more broadly is somewhat limited.

This wide variation in data distributions and model performance points to the larger challenge of generalizability in eHealth studies. To address this issue in future work on eHealth attrition prediction using machine learning, we recommend researchers to (a) consider what aspects of our proposed pipeline may be relevant to their specific context and (b) incorporate more advanced transfer learning techniques. Transfer learning is a machine learning method that leverages knowledge learned from one problem and applies it to a related but different problem. Advanced transfer learning techniques can enhance DMHIs by using existing knowledge, addressing class imbalance and feature extraction, and incorporating insights from large datasets to drive actionable solutions for reducing attrition and increasing engagement in DMHIs.

Table 4. Sensitivity of attrition prediction model performance to imputation. Ablated versions of the proposed pipeline without imputing missing values are compared to the full pipeline in terms of f_1, area under curve, and accuracy scores. All models used extreme gradient boosting (XGBoost) and trained and were tested on all feature sets of the Managing Anxiety (MA), Future Thinking (FT), and Calm Thinking (CT) studies.

Data	XGBoost Model Version	Evaluation Metric		
		F1-macro↑	Area Under Curve↑	Accuracy↑
MA		146 features		
	No Imputation	0.715	0.801	0.716
	Imputation	**0.832**	**0.848**	**0.858**
FT		87 features		
	No Imputation	0.726	0.796	0.729
	Imputation	**0.770**	**0.844**	**0.771**
CT		127 features		
	No Imputation	0.905	0.904	0.910
	Imputation	**0.917**	**0.926**	**0.918**

Note. ↑ indicates that higher values are more desirable for a given metric (which each can range from 0 to 1, with f_1 and area under curve values above .5 and accuracy values above .7 generally considered reasonable; see section Evaluation Metrics for details). The highest values for each metric are in boldface.

Applied Example

Low engagement in a DMHI may manifest as low initial uptake, substantial early dropout, or failure to adhere long term to the intervention techniques intended to change behavior. Predicting attrition is complicated by the many reasons a person may drop out (e.g., the program is not meeting their needs or already has met their needs). Still, identifying participants at high risk for dropout at an early stage may enable allocation of further support specifically to users who may need it, thus improving engagement while retaining scalability [92]. For example, we implemented a probability prediction algorithm in the CT study (instead of the binary classification algorithm used in the present paper) to predict each participant's probability of not completing the second session. This probability, the user's *attrition risk score* (ARS), was then compared with a threshold τ set by the project coordinator (based on a goal to have roughly equal cell sizes after the second randomization point in the study's SMART design). Participants (n = 547) for whom ARS $\geq \tau$ were deemed to have a higher risk of dropping out and were then randomized to receive supplemental telecoaching (n = 282) or not (n = 265). Those in the coaching condition received an email connecting them with their coach, who proposed a phone call to discuss study goals, reinforce use, and address any technical issues or other study questions. We excluded higher-risk participants randomized to supplemental telecoaching (n = 282) from analyses for the present paper. For more details about this implementation, see the CT main outcomes paper [30].

Limitations

One limitation of our analyses is that we focused on participants who started Session 1 training and excluded many participants who dropped out before that point. Another limitation is that we had to use the existing features of the studies, which narrowed our options for feature extraction. It is possible that the model would be further improved with more detailed features (e.g., user continuous location [GPS]; passive detection of more finely grained user behavior at the level of individual items vs. at the level of scale scores or the entire training or assessment task). Additionally, the feature importance results should be interpreted cautiously; readers should refrain from inferring a causal relationship between these features and early attrition. Further research is needed to establish the extent to which such features cause or are a consequence of risk for attrition; it might also be informative to evaluate different imputation and modeling strategies. Furthermore, we employed imputation strategies for all missing numeric values, even in cases where dropout meant the meaning of a given measure no longer applied (e.g., for Return Intention, imputing number of days expected to return for Session 2 even when the participant did not complete Session 1; for Impact of Anxious Imagery Prime, imputing peak anxiety during the prime even when the participant started training but never completed the prime). Future studies should consider (a) removing features containing missing values that cannot be meaningfully imputed or (b) restricting the sample to participants who completed all features that cannot be meaningfully imputed. Finally, future work should seek to identify and, if needed, mitigate potential algorithmic biases. For example, the present studies required participants to have internet access and were optimized for computer delivery, which may lead to underrepresentation in the training data for demographic groups that lack internet access or are dependent on smartphones [120, 121]. While some studies have shown that including demographic features (e.g., gender, race) in early dropout prediction has minimal impact on algorithmic fairness [112], it is prudent to perform a sensitivity analysis excluding these features, to compare model performance by demographic group, and to employ bias-aware model calibration techniques when possible [113].

CONCLUSION

The present analyses aimed to identify participants at high risk of dropout during the early stage of three multi-session web-based CBM-I studies using a combination of self-reported and passively detected measures. Our findings suggest that features involving passive detection of user behavior, such as time spent on a web page of CBM-I training exercises, time of day, latency in completing assessments, and the type of device used, were the most informative predictors of attrition. Additionally, our results showed that using all features extracted from a given study led to the best predictive performance, highlighting the importance of using a combination of feature types when predicting attrition. Consequently, using passive indicators of user behavior, in conjunction with self-reported measures, can increase the accuracy of predicting dropout in web-based CBM-I studies. Although our pipeline provides a framework to consider when predicting attrition in DMHIs, many interesting, open questions remain about how extensively our findings generalize to other CBM-I studies (e.g., in populations with diagnosed anxiety [vs. trait anxiety]; in mobile app-based [vs. web-based] CBM-I studies; in CBM-Is embedded in managed care settings [vs. on a public website]) and to DMHIs more broadly (e.g., unguided web-based cognitive behavioral therapy; [122]). Our analyses highlight the challenge of generalizability in DMHI studies and the need for more personalized attrition prevention strategies. Overall, our results emphasize the potential value of understanding user behavior in early stages of the program and using it as a predictor of dropout, which may guide development of more effective and efficient DMHIs.

ACKNOWLEDGMENTS

This work was supported in part by NIMH R01MH113752, NIMH R34MH106770, NIMH R01MH132138 and a Templeton Science of Prospection Research Award. The authors thank the Sensing Systems for Health (S^2He) Lab; Program for Anxiety, Cognition, and Treatment; and the MindTrails team at the University of Virginia for their feedback and work developing the MindTrails platform.

REFERENCES

[1] Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. Bioinformatics, 26(10), 1340–1347.

[2] Baumel, A., Muench, F., Edan, S., Kane, J. M., et al. (2019). Objective user engagement with mental health apps: systematic search and panel-based usage analysis. Journal of Medical Internet Research, 21(9), e14567.

[3] Beck, A. T. (1979). Cognitive therapy and the emotional disorders. Penguin.

[4] Beck, A. T., & Clark, D. A. (1997). An information processing model of anxiety: Automatic and strategic processes. Behaviour Research and Therapy, 35(1), 49–58.

[5] Bellinger, C., Sharma, S., & Japkowicz, N. (2012). One-class versus binary classification: Which and when?. In 2012 11th International Conference on Machine Learning and Applications, Vol. 2. IEEE, 102–106.

[6] Berger, T., Hämmerli, K., Gubser, N., Andersson, G., & Caspar, F. (2011). Internet-based treatment of depression: a randomized controlled trial comparing guided with unguided self-help. Cognitive Behaviour Therapy, 40(4), 251–266.

[7] Bergstra, J., Yamins, D., Cox, D. D., et al. (2013). Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In Proceedings of the 12th Python in Science Conference, Vol. 13. Citeseer, 20.

[8] Berry, N., Lobban, F., & Bucci, S. (2019). A qualitative exploration of service user views about using digital health interventions for self-management in severe mental health problems. BMC Psychiatry, 19(1), 1–13.

[9] Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A. (2021). A review on outlier/anomaly detection in time series data. ACM Computing Surveys (CSUR), 54(3), 1–33.

[10] Borghouts, J., Eikey, E., Mark, G., De Leon, C., Schueller, S. M., Schneider, M., Stadnick, N., Zheng, K., Mukamel, D., & Sorkin, D. H. (2021). Barriers to and Facilitators of User Engagement With Digital Mental Health Interventions: Systematic Review. Journal of Medical Internet Research, 23(3), e24387.

[11] Borkovec, T. D., & Nau, S. D. (1972). Credibility of analogue therapy rationales. Journal of Behavior Therapy and Experimental Psychiatry, 3(4), 257–260.

[12] Carleton, R. N., Norton, P. J., & Asmundson, G. J. G. (2007). Fearing the unknown: A short version of the Intolerance of Uncertainty Scale. Journal of Anxiety Disorders, 21(1), 105–117.

[13] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.

[14] Chen, G., Gully, S. M., & Eden, D. (2001). Validation of a new general self-efficacy scale. Organizational Research Methods, 4(1), 62–83.

[15] Chen, T. (2014, October 22). Introduction to boosted trees [Slides]. University of Washington Computer Science. Retrieved from https://web.njit.edu/~usman/courses/cs675_fall16/BoostedTree.pdf

[16] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., et al. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4).

[17] Cheung, F., & Lucas, R. E. (2014). Assessing the validity of single-item life satisfaction measures: Results from three large samples. Quality of Life Research, 23(10), 2809–2818.

[18] Cheung, K., Ling, W., Karr, C. J., Weingardt, K., Schueller, S. M., & Mohr, D. C. (2018). Evaluation of a recommender app for apps for the treatment of depression and anxiety: an analysis of longitudinal user engagement. Journal of the American Medical Informatics Association, 25(8), 955–962.

[19] Chien, I., Enrique, A., Palacios, J., Regan, T., Keegan, D., Carter, D., Tschiatschek, S., Nori, A., Thieme, A., Richards, D., Doherty, G., & Belgrave, D. (2020). A Machine Learning Approach to Understanding Patterns of Engagement With Internet-Delivered Mental Health Interventions. JAMA Network Open, 3(7), e2010791–e2010791.

[20] Clark, D. M., Salkovskis, P. M., Öst, L.-G., Breitholtz, E., Koehler, K. A., Westling, B. E., Jeavons, A., & Gelder, M. (1997). Misinterpretation of body sensations in panic disorder. Journal of Consulting and Clinical Psychology, 65(2), 203.

[21] Collins, R. L., Parks, G. A., & Marlatt, G. A. (1985). Social determinants of alcohol consumption: the effects of social interaction and model status on the self-administration of alcohol. Journal of Consulting and Clinical Psychology, 53(2), 189.

[22] Crafoord, M.-T., Fjell, M., Sundberg, K., Nilsson, M., & Langius-Eklöf, A. (2020). Engagement in an Interactive App for Symptom Self-Management during Treatment in Patients With Breast or Prostate Cancer: Mixed Methods Study. Journal of Medical Internet Research, 22(8), e17058.

[23] Daniel, K. E., Eberle, J. W., & Teachman, B. A. (2020). Web-Based Interpretation Bias Training to Reduce Anxiety: A Sequential, Multiple-Assignment, Randomized Trial. Retrieved from https://doi.org/10.17605/OSF.IO/AF4N7 Preregistration.

[24] Dennis, J. P., & Vander Wal, J. S. (2010). The cognitive flexibility inventory: Instrument development and estimates of reliability and validity. Cognitive Therapy and Research, 34(3), 241–253.

[25] Dixon, L. J., & Linardon, J. (2019). A systematic review and meta-analysis of dropout rates from dialectical behaviour therapy in randomized controlled trials. Cognitive Behaviour Therapy, 1–16.

[26] Doherty, K., & Doherty, G. (2018). Engagement in HCI: conception, theory and measurement. ACM Computing Surveys (CSUR), 51(5), 1–39.

[27] Dweck, C. S. (2006). Mindset: The new psychology of success. Random House. New York, NY.

[28] Eberle, J. W., Baee, S., Behan, H. C., Baglione, A. N., Boukhechba, M., Funk, D. H., Barnes, L. E., & Teachman, B. A. (2022). TeachmanLab/MT-Data-CalmThinkingStudy: Centralized Data Cleaning for MindTrails Calm Thinking Study. Retrieved from https://doi.org/10.5281/zenodo.6192907
[29] Eberle, J. W., Boukhechba, M., Sun, J., Zhang, D., Funk, D., Barnes, L., & Teachman, B. (2023). Shifting Episodic Prediction With Online Cognitive Bias Modification: A Randomized Controlled Trial. Clinical Psychological Science. Retrieved from https://doi.org/10.1177/21677026221103128
[30] Eberle, J. W., Daniel, K. E., Baee, S., Silverman, A. L., Lewis, E., Baglione, A. N., Werntz, A., French, N. J., Ji, J. L., Hohensee, N., Tong, X., Huband, J. M., Boukhechba, M., Funk, D. H., Barnes, L. E., & Teachman, B. A. (2024). Web-based interpretation bias training to reduce anxiety: A sequential multiple-assignment randomized trial. PsyArXiv. https://doi.org/10.31234/osf.io/37k8n
[31] Elreedy, D., & Atiya, A. F. (2019). A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. Information Sciences, 505, 32–64.
[32] Ersner-Hershfield, H., Garton, M. T., Ballard, K., Samanez-Larkin, G. R., & Knutson, B. (2009). Don't stop thinking about tomorrow: Individual differences in future self-continuity account for saving. Judgment and Decision making, 4(4), 280.
[33] Eysenbach, G. (2005). The law of attrition. Journal of Medical Internet Research, 7(1), e11.
[34] Fernandez, E., Salem, D., Swift, J. K., & Ramtahal, N. (2015). Meta-analysis of dropout from cognitive behavioral therapy: Magnitude, timing, and moderators. Journal of Consulting and Clinical Psychology, 83(6), 1108.
[35] Flanagan, J. C. (1978). A research approach to improving our quality of life. American Psychologist, 33(2), 138.
[36] Fodor, L. A., Georgescu, R., Cuijpers, P., Szamoskozi, Ş., David, D., Furukawa, T. A., & Cristea, I. A. (2020). Efficacy of cognitive bias modification interventions in anxiety and depressive disorders: a systematic review and network meta-analysis. The Lancet Psychiatry, 7(6), 506–514.
[37] Francis, A. W., Dawson, D. L., & Golijani-Moghaddam, N. (2016). The development and validation of the Comprehensive assessment of Acceptance and Commitment Therapy processes (CompACT). Journal of Contextual Behavioral Science, 5(3), 134–145.
[38] Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmospheric Environment, 32(14-15), 2627–2636.
[39] Gersh, E., Hallford, D. J., Rice, S. M., Kazantzis, N., Gersh, H., Gersh, B., & McCarty, C. A. (2017). Systematic review and meta-analysis of dropout rates in individual psychotherapy for generalized anxiety disorder. Journal of Anxiety Disorders, 52, 25–33.
[40] Goldberg, S. B., Bolt, D. M., & Davidson, R. J. (2021). Data Missing Not at Random in Mobile Health Research: Assessment of the Problem and a Case for Sensitivity Analyses. Journal of Medical Internet Research, 23.
[41] Goyal, S., Morita, P., Lewis, G. F., Yu, C., Seto, E., & Cafazzo, J. A. (2016). The systematic design of a behavioural mobile health application for the self-management of type 2 diabetes. Canadian Journal of Diabetes, 40(1), 95–104.
[42] Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. Journal of Personality and Social Psychology, 85(2), 348.
[43] Hallion, L. S., & Ruscio, A. M. (2011). A meta-analysis of the effect of cognitive bias modification on anxiety and depression. Psychological Bulletin, 137(6), 940.
[44] Harari, G. M. (2020). A process-oriented approach to respecting privacy in the context of mobile phone tracking. Current Opinion in Psychology, 31, 141–147.
[45] Henkemans, O. A. B., Rogers, W. A., & Dumay, A. (2011). Personal characteristics and the law of attrition in randomized controlled trials of eHealth services for self-care. Gerontechnology.
[46] Hohensee, N., Meyer, M. J., & Teachman, B. A. (2020). The effect of confidence on dropout rate and outcomes in online cognitive bias modification. Journal of Technology in Behavioral Science, 1–9.
[47] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.
[48] Hutchesson, M. J., Duncan, M. J., Oftedal, S., Ashton, L. M., Oldmeadow, C., Kay-Lambkin, F., & Whatnall, M. C. (2021). Latent class analysis of multiple health risk behaviors among Australian university students and associations with psychological distress. Nutrients, 13(2), 425.
[49] Ji, J. L., Baee, S., Zhang, D., Calicho-Mamani, C. P., Meyer, M. J., Funk, D., ... & Teachman, B. A. (2021). Multi-session online interpretation bias training for anxiety in a community sample. Behaviour Research and Therapy, 103864.
[50] Johnson, R., & Zhang, T. (2013). Learning nonlinear functions using regularized greedy forest. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(5), 942–954.
[51] Jones, E. B., & Sharpe, L. (2017). Cognitive bias modification: A review of meta-analyses. Journal of Affective Disorders, 223, 175–183.
[52] Kim, K.-Y., Kim, B.-J., & Yi, G.-S. (2004). Reuse of imputed data in microarray analysis increases imputation efficiency. BMC Bioinformatics, 5(1), 1–9.
[53] Kim, S., Choi, D., Lee, E., & Rhee, W. (2017). Churn prediction of mobile and online casual games using play log data. PLOS ONE, 12(7), e0180735.
[54] Kovacs, G., Wu, Z., & Bernstein, M. S. (2018). Rotating Online Behavior Change Interventions Increases Effectiveness But Also Increases Attrition. Proceedings of the ACM on Human-Computer Interaction, 2(CSCW), 1–25.
[55] Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2003). The Patient Health Questionnaire-2: validity of a two-item depression screener. Medical Care.
[56] Kroenke, K., Spitzer, R. L., Williams, J. B. W., Monahan, P. O., & Löwe, B. (2007). Anxiety disorders in primary care:

prevalence, impairment, comorbidity, and detection. Annals of Internal Medicine, 146(5), 317–325.
[57] Kumari, R., & Srivastava, S. K. (2017). Machine learning: A review on binary classification. International Journal of Computer Applications, 160(7).
[58] Kwon, H., Kim, H. H., An, J., Lee, J.-H., & Park, Y. R. (2021). Lifelog Data-Based Prediction Model of Digital Health Care App Customer Churn: Retrospective Observational Study. Journal of Medical Internet Research, 23(1), e22184.
[59] Lattie, E. G., Adkins, E. C., Winquist, N., Stiles-Shields, C., Wafford, Q. E., & Graham, A. K. (2019). Digital Mental Health Interventions for Depression, Anxiety, and Enhancement of Psychological Well-Being Among College Students: Systematic Review. Journal of Medical Internet Research, 21(7), e12869.
[60] Lattie, E. G., Adkins, E. C., Winquist, N., Stiles-Shields, C., Wafford, Q. E., & Graham, A. K. (2019). Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. Journal of Medical Internet Research, 21(7), e12869.
[61] Linardon, J., & Fuller-Tyszkiewicz, M. (2020). Attrition and adherence in smartphone-delivered interventions for mental health problems: A systematic and meta-analytic review. Journal of Consulting and Clinical Psychology, 88(1), 1.
[62] Linardon, J., Shatte, A., Tepper, H., & Fuller-Tyszkiewicz, M. (2020). A survey study of attitudes toward, and preferences for, e-therapy interventions for eating disorder psychopathology. International Journal of Eating Disorders.
[63] Lipschitz, J. M., Van Boxtel, R., Torous, J., Firth, J., Lebovitz, J. G., Burdick, R., ... & Hogan, T. P. (2022). Digital mental health interventions for depression: Scoping review of user engagement. J Med Internet Res, 24(10).
[64] Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. Behaviour Research and Therapy, 33(3), 335–343.
[65] MacLeod, C., & Mathews, A. (2012). Cognitive bias modification approaches to anxiety. Annual Review of Clinical Psychology, 8, 189–217.
[66] Pujianto, U., Wibawa, A. P., & Akbar, M. I. (2019, October). K-nearest neighbor (k-NN) based missing data imputation. In 2019 5th International Conference on Science in Information Technology (ICSITech) (pp. 83-88). IEEE.
[67] Martinez, A. B., Lau, J., Brown, J. S., et al. (2020). Filipino help-seeking for mental health problems and associated barriers and facilitators: a systematic review. Social Psychiatry and Psychiatric Epidemiology.
[68] Mathews, A., & Mackintosh, B. (1998). A cognitive model of selective processing in anxiety. Cognitive Therapy and Research, 22(6), 539–560.
[69] Mathews, A., & Mackintosh, B. (2000). Induced emotional interpretation bias and anxiety. Journal of Abnormal Psychology, 109(4), 602.
[70] Miyamoto, S., Dharmar, M., Fazio, S., Tang-Feldman, Y., & Young, H. M. (2018). mHealth Technology and Nurse Health Coaching to Improve Health in Diabetes: Protocol for a Randomized Controlled Trial. JMIR Research Protocols, 7(2), e45.
[71] Mohr, D. C., Hart, S. L., Howard, I., Julian, L., Vella, L., Catledge, C., & Feldman, M. D. (2006). Barriers to psychotherapy among depressed and nondepressed primary care patients. Annals of Behavioral Medicine, 32(3), 254–258.
[72] Mohr, D. C., Ho, J., Duffecy, J., Baron, K. G., Lehman, K. A., Jin, L., & Reifler, D. (2010). Perceived barriers to psychological treatments and their relationship to depression. Journal of Clinical Psychology, 66(4), 394–409.
[73] Mowbray, F. I., Fox-Wasylyshyn, S. M., & El-Masri, M. M. (2019). Univariate outliers: a conceptual overview for the nurse researcher. Canadian Journal of Nursing Research, 51(1), 31–37.
[74] Murray, E., White, I. R., Varagunam, M., Godfrey, C., Khadjesari, Z., & McCambridge, J. (2013). Attrition revisited: adherence and retention in a web-based alcohol trial. Journal of Medical Internet Research, 15(8), e162.
[75] Nahum-Shani, I., Shaw, S. D., Carpenter, S. M., Murphy, S. A., & Yoon, C. (2022). Engagement in digital interventions. American Psychologist.
[76] Namaky, N., Glenn, J. J., Eberle, J. W., & Teachman, B. A. (2021). Adapting cognitive bias modification to train healthy prospection. Behaviour Research and Therapy, 144, 103923.
[77] Nembrini, S., König, I. R., & Wright, M. N. (2018). The revival of the Gini importance? Bioinformatics, 34(21), 3711–3718.
[78] Newby, K., Teah, G., Cooke, R., Li, X., Brown, K., Salisbury-Finch, B., ... & Allott, K. (2021). Do automated digital health behaviour change interventions have a positive effect on self-efficacy? A systematic review and meta-analysis. Health Psychology Review, 15(1), 140–158.
[79] Nicholas, J., Proudfoot, J., Parker, G., Gillis, I., Burckhardt, R., Manicavasagar, V., ... & Hadzi-Pavlovic, D. (2010). The Ins and Outs of an Online Bipolar Education Program: A Study of Program Attrition. Journal of Medical Internet Research, 12(5), e57.
[80] Nicholas, J., Proudfoot, J., Parker, G., Gillis, I., Burckhardt, R., Manicavasagar, V., Smith, M., et al. (2010). The ins and outs of an online bipolar education program: a study of program attrition. Journal of Medical Internet Research, 12(5), e1450.
[81] Noble, W. S. (2006). What is a support vector machine? Nature Biotechnology, 24(12), 1565–1567.
[82] Norman, S. B., Cissell, S. H., Means-Christensen, A. J., & Stein, M. B. (2006). Development and validation of an overall anxiety severity and impairment scale (OASIS). Depression and Anxiety, 23(4), 245–249.
[83] Perianez, A., Saas, A., Guitart, A., & Magne, C. (2016). Churn Prediction in Mobile Social Games: Towards a Complete Assessment Using Survival Ensembles. In 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA) (pp. 564–573). IEEE.

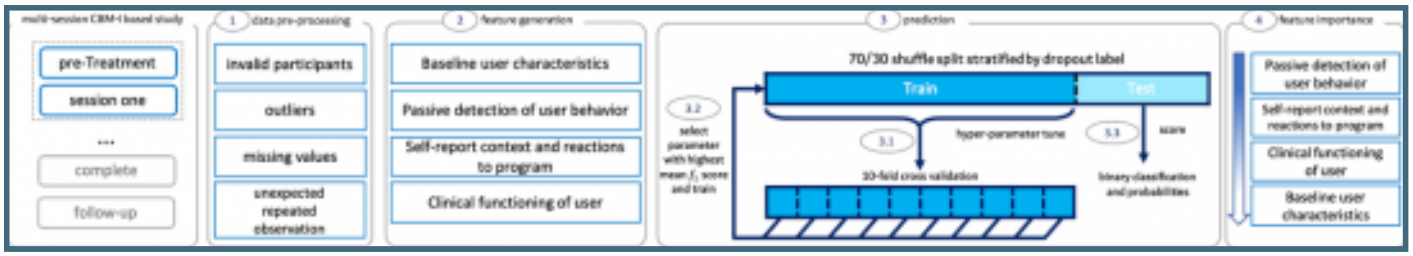
[84] Pratap, A., Chaibub Neto, E., Snyder, P., Stepnowsky, C., Elhadad, N., Grant, D., ... & Wilbanks, J. (2020). Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants. NPJ Digital Medicine, 3(1), 1–10.
[85] Price, M., Gros, D. F., McCauley, J. L., Gros, K. S., & Ruggiero, K. J. (2012). Nonuse and dropout attrition for a web-based mental health intervention delivered in a post-disaster context. Psychiatry: Interpersonal and Biological Processes, 75(3), 267–284.
[86] Rabbi, M., Kotov, M. P., Cunningham, R., Bonar, E. E., Nahum-Shani, I., Klasnja, P., ... & Murphy, S. (2018). Toward increasing engagement in substance use data collection: development of the Substance Abuse Research Assistant app and protocol for a microrandomized trial using adolescents and emerging adults. JMIR Research Protocols, 7(7), e166.
[87] Salemink, E., Kindt, M., Rienties, H., & Van Den Hout, M. (2014). Internet-based cognitive bias modification of interpretations in patients with anxiety disorders: a randomised controlled trial. Journal of Behavior Therapy and Experimental Psychiatry, 45(1), 186–195.
[88] Santomauro, D. F., Herrera, A. M. M., Shadid, J., Zheng, P., Ashbaugh, C., Pigott, D. M., ... & Murray, C. J. (2021). Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. The Lancet, 398(10312), 1700–1712.
[89] Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): a reevaluation of the Life Orientation Test. Journal of Personality and Social Psychology, 67(6), 1063.
[90] Scherer, E. A., Ben-Zeev, D., Li, Z., & Kane, J. M. (2017). Analyzing mHealth engagement: joint models for intensively collected user engagement data. JMIR mHealth and uHealth, 5(1), e6474.
[91] Schroé, H., Crombez, G., De Bourdeaudhuij, I., Van Dyck, D., et al. (2022). Investigating When, Which, and Why Users Stop Using a Digital Health Intervention to Promote an Active Lifestyle: Secondary Analysis With A Focus on Health Action Process Approach–Based Psychological Determinants. JMIR mHealth and uHealth, 10(1), e30583.
[92] Schueller, S. M., Tomasino, K. N., & Mohr, D. C. (2017). Integrating Human Support Into Behavioral Intervention Technologies: The Efficiency Model of Support. Clinical Psychology: Science and Practice, 24(1), 27–45.
[93] Schure, M. B., Lindow, J. C., Greist, J. H., Nakonezny, P. A., Bailey, S. J., Bryan, W. L., & Byerly, M. J. (2019). Use of a Fully Automated Internet-Based Cognitive Behavior Therapy Intervention in a Community Population of Adults With Depression Symptoms: Randomized Controlled Trial. Journal of Medical Internet Research, 21(11), e14754.
[94] Stieger, M., Flückiger, C., Rüegger, D., Kowatsch, T., Roberts, B. W., & Allemand, M. (2021). Changing personality traits with the help of a digital personality change intervention. Proceedings of the National Academy of Sciences, 118(8).
[95] Stiles-Shields, C., Montague, E., Lattie, E. G., Kwasny, M. J., & Mohr, D. C. (2017). What might get in the way: barriers to the use of apps for depression. Digital Health, 3, 2055207617713827.
[96] De, S., & Prabu, P. (2022). Predicting customer churn: A systematic literature review. Journal of Discrete Mathematical Sciences and Cryptography, 25(7), 1965-1985.
[97] Tahsin, F., Tracy, S., Chau, E., Harvey, S., Loganathan, M., McKinstry, B., ... & Mercer, S. W. (2021). Exploring the relationship between the usability of a goal-oriented mobile health application and non-usage attrition in patients with multimorbidity: A blended data analysis approach. Digital Health, 7, 20552076211045579.
[98] Taki, S., Lymer, S., Russell, C. G., Campbell, K., Laws, R., Ong, K. L., ... & Wen, L. M. (2017). Assessing user engagement of an mHealth intervention: development and implementation of the growing healthy app engagement index. JMIR mHealth and uHealth, 5(6), e7236.
[99] Tang, X., Liu, Y., Shah, N., Shi, X., Mitra, P., & Wang, S. (2020). Knowing your FATE: Friendship, Action and Temporal Explanations for User Engagement Prediction on Social Apps. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2269–2279). ACM.
[100] Torous, J., Lipschitz, J., Ng, M., & Firth, J. (2019). Dropout rates in clinical trials of smartphone apps for depressive symptoms: a systematic review and meta-analysis. Journal of Affective Disorders.
[101] Twomey, C., O'Reilly, G., Byrne, M., Bury, M., White, A., Kissane, S., ... & Clancy, N. (2014). A randomized controlled trial of the computerized CBT programme, MoodGYM, for public mental health service users waiting for interventions. British Journal of Clinical Psychology, 53(4), 433–450.
[102] Case Western Reserve University (2010). Readiness Ruler. Retrieved from https://case.edu/socialwork/centerforebp/resources/readiness-ruler
[103] Werntz, A., Silverman, A. L., Behan, H., Patel, S. K., Beltzer, M., Boukhechba, M. O., ... & Teachman, B. A. (2022). Lessons Learned: Providing Supportive Accountability in an Online Anxiety Intervention. Behavior Therapy, 53(3), 492–507.
[104] Wolpe, J. (1969). The practice of behavior therapy. Pergamon. https://psycnet.apa.org/record/1970-18837-000Wong, H. W., Lo, B., Shi, J., Hollenberg, E., Abi-Jaoude, A., Johnson, A., ... & Henderson, J. (2021). Postsecondary student engagement with a mental health app and online platform (Thought Spot): qualitative study assessing factors related to user experience. JMIR Mental Health, 8(4), e23447.
[105] Wong, H. W., Lo, B., Shi, J., Hollenberg, E., Abi-Jaoude, A., Johnson, A., ... & Henderson, J. (2021). Postsecondary student engagement with a mental health app and online platform (Thought Spot): qualitative study assessing factors related to user experience. JMIR Mental Health, 8(4), e23447.
[106] Xie, L. F., Itzkovitz, A., Roy-Fleming, A., Da Costa, D., & Brazeau, A. S. (2020). Understanding Self-Guided Web-Based Educational Interventions for Patients With Chronic Health Conditions: Systematic Review of Intervention Features and

Adherence. Journal of Medical Internet Research, 22(8), e18355.
[107] Yang, C., Shi, X., Jie, L., & Han, J. (2018). I Know You'll Be Back: Interpretable New User Clustering and Churn Prediction on a Mobile Social Application. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 914–922). ACM.
[108] Yang, R., Cui, L., Li, F., Xiao, J., Zhang, Q., & Oei, T. P. S. (2017). Effects of cognitive bias modification training via smartphones. Frontiers in Psychology, 8, 1370.
[109] Yang, S., Zhou, P., Duan, K., Hossain, M. S., & Alhamid, M. F. (2018). emHealth: towards emotion health through depression prediction and intelligent health recommender system. Mobile Networks and Applications, 23(2), 216–226.
[110] Yardley, L., Choudhury, T., Patrick, K., & Michie, S. (2016). Current issues and future directions for research into digital behavior change interventions. American Journal of Preventive Medicine, 51(5), 814–815.
[111] Jardine, J., Nadal, C., Robinson, S., Enrique, A., Hanratty, M., & Doherty, G. (2023). Between Rhetoric and Reality: Real-world Barriers to Uptake and Early Engagement in Digital Mental Health Interventions. ACM Transactions on Computer-Human Interaction.
[112] Yu, R., Lee, H., & Kizilcec, R. F. (2021, June). Should college dropout prediction models include protected attributes?. In Proceedings of the Eighth ACM Conference on Learning@Scale (pp. 91-100).
[113] Karimi-Haghighi, M., Castillo, C., Hernandez-Leo, D., & Oliver, V. M. (2021). Predicting early dropout: Calibration and algorithmic fairness considerations. arXiv preprint arXiv:2103.09068.
[114] Goldberg, Simon B., Daniel M. Bolt, and Richard J. Davidson. "Data missing not at random in mobile health research: assessment of the problem and a case for sensitivity analyses." Journal of medical Internet research 23.6 (2021): e26749.
[115] Linardon, J., & Fuller-Tyszkiewicz, M. (2020). Attrition and adherence in smartphone-delivered interventions for mental health problems: A systematic and meta-analytic review. Journal of consulting and clinical psychology, 88(1), 1.
[116] Kazdin, A. E., & Rabbitt, S. M. (2013). Novel models for delivering mental health services and reducing the burdens of mental illness. Clinical Psychological Science, 1(2), 170-191.
[117] Muñoz, A. O., Camacho, E., & Torous, J. (2021). Marketplace and literature review of Spanish language mental health apps. Frontiers in Digital Health, 3, 615366.
[118] Stringer, H. (2024, January 1). Mental health care is in high demand. Psychologists are leveraging tech and peers to meet the need. Monitor on Psychology, 55(1). https://www.apa.org/monitor/2024/01/trends-pathways-access-mental-health-care
[119] Schueller, S. M., Tomasino, K. N., & Mohr, D. C. (2017). Integrating human support into behavioral intervention technologies: The efficiency model of support. Clinical Psychology: Science and Practice, 24(1), 27.
[120] Lorence, D. P., Park, H., & Fox, S. (2006). Racial disparities in health information access: resilience of the digital divide. Journal of medical systems, 30, 241-249.
[121] Tsetsi, E., & Rains, S. A. (2017). Smartphone Internet access and use: Extending the digital divide and usage gap. Mobile Media & Communication, 5(3), 239-255.
[122] Morgan, C., Mason, E., Newby, J. M., Mahoney, A. E., Hobbs, M. J., McAloon, J., & Andrews, G. (2017). The effectiveness of unguided internet cognitive behavioural therapy for mixed anxiety and depression. Internet interventions, 10, 47-53.
[123] Schueller, S. M., Washburn, J. J., & Price, M. (2016). Exploring mental health providers' interest in using web and mobile-based tools in their practices. Internet interventions, 4, 145-151.
[124] Berry, N., Bucci, S., & Lobban, F. (2017). Use of the internet and mobile phones for self-management of severe mental health problems: qualitative study of staff views. JMIR mental health, 4(4), e8311.
[125] Schueller, S. M., Washburn, J. J., & Price, M. (2016). Exploring mental health providers' interest in using web and mobile-based tools in their practices. Internet interventions, 4, 145-151.
[126] van Bronswijk, S. C., DeRubeis, R. J., Lemmens, L. H., Peeters, F. P., Keefe, J. R., Cohen, Z. D., & Huibers, M. J. (2021). Precision medicine for long-term depression outcomes using the Personalized Advantage Index approach: cognitive therapy or interpersonal psychotherapy?. Psychological medicine, 51(2), 279-289.

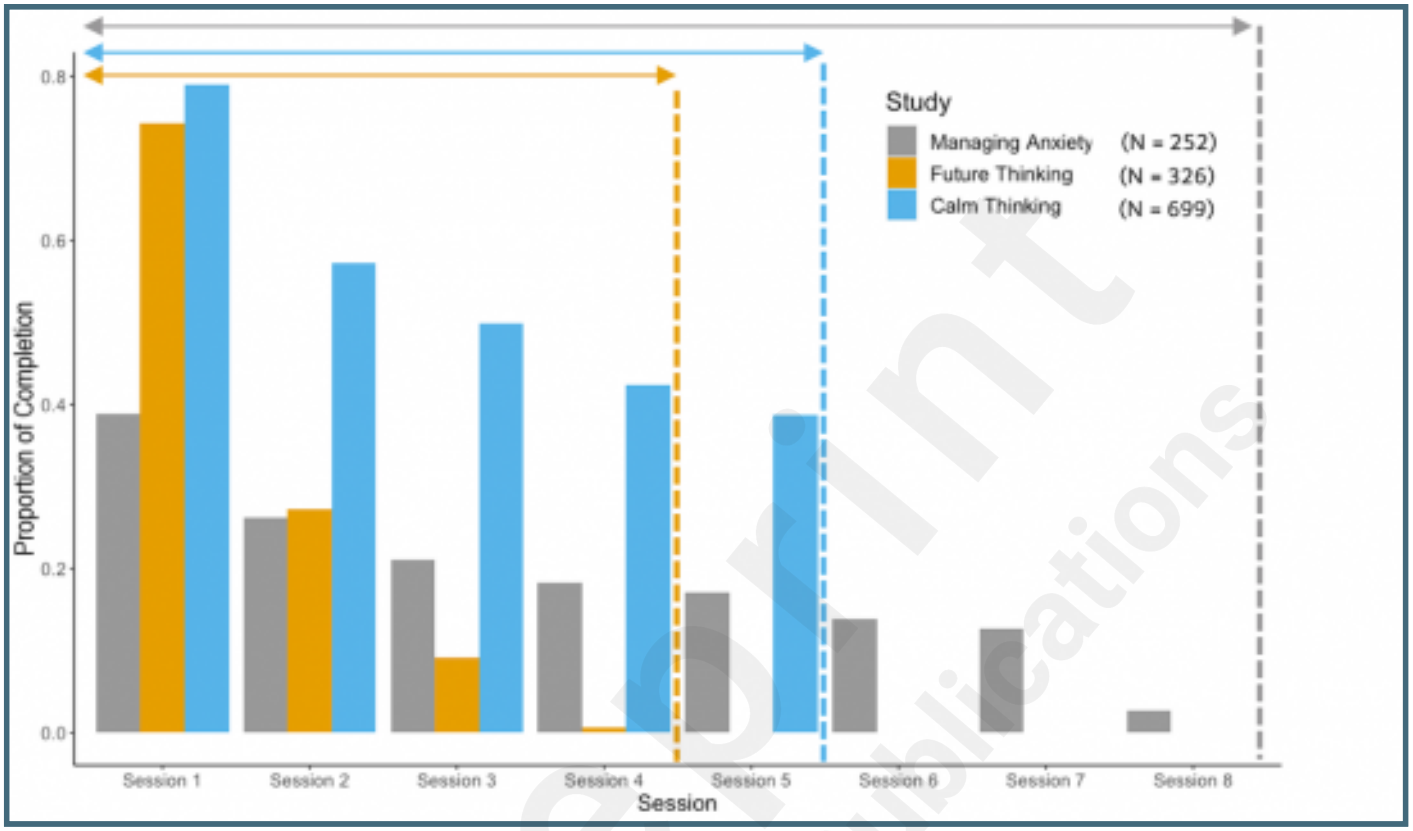
Supplementary Files

Figures

Overview of pipeline predicting early-stage attrition in web-based multi-session cognitive bias modifi- cation for interpretation (CBM-I) interventions.



Proportion of completion per training session (out of participants who started Session 1 training) by study. The session was deemed completed if participants completed the last questionnaire in the assessment that immediately followed the training session. Dashed lines show the last training session for each study.



Importance level of each feature set relative to other feature sets for early attrition prediction in CBM-I studies. Gini importance scores averaged across two iterations are shown. We use XGBoost classifier since it performed the best. These scores reflect the importance of a feature set relative to others (not absolute importance) and can range from 0 to 1, with higher scores reflecting greater importance. Horizontal bars reflect the median score; dots represent outliers, which are observations that fall outside of the boxplot; and whiskers represent the minimum and maximum observations within 1.5 times the interquartile range from the lower and upper quartiles, respectively. Note: No important baseline user characteristic features emerged for the Calm Thinking study.