PLOS ONE

RESEARCH ARTICLE

Deep learning for the prediction of clinical outcomes in internet-delivered CBT for depression and anxiety

Niranjani Prasad[1]*, Isabel Chien[2], Tim Regan[3], Angel Enrique[4,5], Jorge Palacios[4,5], Dessie Keegan[4], Usman Munir[1], Ryutaro Tanno[6], Hannah Richardson[1], Aditya Nori[1], Derek Richards[4,5‡], Gavin Doherty[4,7‡], Danielle Belgrave[6‡], Anja Thieme[1‡]

1 Microsoft Health Futures, Microsoft Research, Cambridge, United Kingdom, **2** Cambridge University, Cambridge, United Kingdom, **3** Cambridge Respiratory Innovations, Cambridge, United Kingdom, **4** SilverCloud Science, SilverCloud Health, Dublin, Ireland, **5** E-Mental Health Group, School of Psychology, Trinity College Dublin, Dublin, Ireland, **6** GSK, London, United Kingdom, **7** School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland

‡ DR, GD, DB and AT are joint last authors on this work.
* niranjani.prasad@microsoft.com

Abstract

In treating depression and anxiety, just over half of all clients respond. Monitoring and obtaining early client feedback can allow for rapidly adapted treatment delivery and improve outcomes. This study seeks to develop a state-of-the-art deep-learning framework for predicting clinical outcomes in internet-delivered Cognitive Behavioural Therapy (iCBT) by leveraging large-scale, high-dimensional time-series data of client-reported mental health symptoms and platform interaction data. We use de-identified data from 45,876 clients on SilverCloud Health, a digital platform for the psychological treatment of depression and anxiety. We train deep recurrent neural network (RNN) models to predict whether a client will show reliable improvement by the end of treatment using clinical measures, interaction data with the iCBT program, or both. Outcomes are based on total improvement in symptoms of depression (Patient Health Questionnaire-9, PHQ-9) and anxiety (Generalized Anxiety Disorder-7, GAD-7), as reported within the iCBT program. Using internal and external datasets, we compare the proposed models against several benchmarks and rigorously evaluate them according to their predictive accuracy, sensitivity, specificity and AUROC over treatment. Our proposed RNN models consistently predict reliable improvement in PHQ-9 and GAD-7, using past clinical measures alone, with above 87% accuracy and 0.89 AUROC after three or more review periods, outperforming all benchmark models. Additional evaluations demonstrate the robustness of the achieved models across (i) different health services; (ii) geographic locations; (iii) iCBT programs, and (iv) client severity subgroups. Results demonstrate the robust performance of dynamic prediction models that can yield clinically helpful prognostic information ready for implementation within iCBT systems to support timely decision-making and treatment adjustments by iCBT clinical supporters towards improved client outcomes.

Citation: Prasad N, Chien I, Regan T, Enrique A, Palacios J, Keegan D, et al. (2023) Deep learning for the prediction of clinical outcomes in internet-delivered CBT for depression and anxiety. PLoS ONE 18(11): e0272685. https://doi.org/10.1371/journal.pone.0272685

Editor: Hugh Cowley, Public Library of Science, UNITED KINGDOM

Received: July 24, 2022

Accepted: November 2, 2023

Published: November 27, 2023

Data Availability Statement: Data cannot be shared publicly because SilverCloud had agreements with its customers to use this data in de-identified form for understanding usage and improving the service, and chose a research partner to work with on this project. The customer agreements, and reasonable expectations of the end users, do not permit general publication or disclosure of the data. Data are available via request at science@amwell.com for researchers who meet the criteria for access to confidential data.

Funding: This project, or any of the authors, did not receive any funding or payments associated to its completion.

Competing interests: AE, JP, DK, and DR are employees of SilverCloud Health. GD is a cofounder of SilverCloud Health and has a minority shareholding in the company. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

Introduction

Depression and anxiety are primary drivers of disability worldwide [1]. Treatment via cognitive behavioural therapy (CBT) has been established through decades of research, and in recent years, the active ingredients in CBT have been disseminated through the internet and proven effective in yielding reliable improvement in clinical symptoms [2–5]. While internet-based CBT (iCBT) offers many unique advantages, in particular flexible access to treatment resources, maintaining user engagement with digitally delivered behavioural interventions remains challenging [6]. The involvement of a clinical supporter, whose role is to encourage and facilitate clients' use of the digital intervention, has been shown to lead to better treatment engagement and outcomes than self-guided therapy [7, 8]. These clinical supporters frequently communicate with their clients via online messages or telephone conversations to facilitate their learning and the application of self-taught mental health management techniques. The specific skills and techniques of CBT treatment, alongside the clinical relationship, seek to address cognitive and behavioral processes to drive change for clients. The acquisition and practical application of these skills by clients is an essential outcome of all CBT-based treatments [9, 10].

Using technology allows supporters to gain information about clients' engagement with the treatment and potentially to intervene based on this information. This is commensurate with the Feedback-Informed Treatment (FIT) paradigm [11] that builds on the routine outcome monitoring (ROM) framework, as continuous monitoring of symptoms during treatment is critical for effective clinical decision-making [12]. Empirical evidence demonstrates that therapists use of FIT reduces the likelihood of patient deterioration and can lead to lower average duration and cost of treatment compared to controls [13–16]. Indeed, guidelines recommend that depression and anxiety treatments should follow through to remission for clients, yet only about 50% of clients reach remission. Feedback allows the supporting clinician to understand if the treatment is progressing as expected, to understand early if the client is benefitting or likely to benefit, and to make any necessary adjustments to improve the likelihood of treatment response. FIT is independent of any one theoretical approach and focuses on the importance of a culture of feedback throughout therapy. In tandem, a focus on deliberate action to improve quality and outcomes has proved salient [11].

In the case of depression and anxiety, just over half of all clients are expected to respond to treatment [17–20]. Insights early in treatment about prospective outcomes can enable clinical supporters of iCBT interventions to engage in more timely and proactive intervention by increasing their level of client support or re-assessing the suitability of the chosen treatment for an individual. These insights can aid decisions on whether a client needs to be stepped up to different care or whether more treatment sessions are needed for a particular client. This allows for more effective distribution of care resources as well as reductions in the negative impact of having a client remain too long on the wrong care pathway: literature has shown that (real-time) feedback to therapists on expected outcomes for a client can prevent symptom deterioration and improve treatment success [15, 21, 22].

Recent years have begun to see the development of machine learning models for risk stratification or prediction of outcomes from treatment where research is at an early stage when using only baseline data. Still, more sound evidence exists when using routine outcome monitoring data [23–25]. To date, within iCBT contexts, models have been built on modest and selective samples (ranging from fewer than 100 to ~2000 users) and have typically involved post-hoc analysis of RCT data or miscellaneous data sources, with mixed results [25–29]. While RCT data has the advantage that it is often less biased than data collected in a naturalistic cohort study, the modest sample sizes restrict the feasibility of high-capacity machine learning methods and limit the scope for robust validation.

In this work, we present the analysis of a large clinical sample of 45,876 mental health clients for predicting clinical outcomes. Our work leverages some unique opportunities afforded by iCBT interventions, which capture real-time data of client interactions with psychotherapy treatment alongside any changes in clinical symptoms over time. This scale of data also enables the application of a deep learning (DL) framework, which has been shown to achieve state-of-the-art performance in many applications with sequential data, including clinical time series [30]. DL methods can scale to large datasets compared to other statistical and machine learning methods. They have proven effective at modelling complex relationships from high-dimensional and unstructured data inputs—such as fine-grained platform interaction data—making them less reliant on significant feature engineering.

Our analysis aims to demonstrate the feasibility of developing robust classification models using DL for the early prediction of treatment outcomes. Furthermore, within the context of an iCBT intervention for the treatment of depression and anxiety, our research investigates: (i) what user data (in what combination) is most informative within such models, (ii) how early within treatment we can achieve predictions robust enough for use in clinical practice, and (iii) how well do the best-performing models generalize to other data populations.

Materials and methods

Data source and study population

This study leverages de-identified clinical measures and iCBT program interaction data from SilverCloud Health. This evidence-based, online self-administered platform delivers low-intensity iCBT alongside feedback from trained clinical supporters. Within the UK, SilverCloud services are predominantly accessed via client referrals from their GP or other healthcare professionals to IAPT (Improving Access to Psychological Therapies), a program that offers talking therapies to adults to help overcome depression and anxiety. IAPT offers treatment within a stepped care service model that ensures that the most effective but least resource-intensive treatment is delivered first, and care is only stepped-up to more intensive face-to-face treatments if required [31]. In this model, clients initially tend to access self-help oriented treatments, which are referred to as 'low intensity' (LI) interventions and that are usually delivered by clinical supporters [17]. SilverCloud Health is the most accessed online LI intervention. It is available in over 80% of IAPT services and offers over 35 treatment programs [32, 33].

Amongst the most popular and most accessed is the "Space from Depression and Anxiety" program. For our data-intense analysis, we considered all clients enrolled in this program between January 2015 and March 2019. The program consists of eight core modules covering fundamental CBT principles for treating symptoms of depression and anxiety. Content is delivered using textual and audio-visual materials, interactive tools (journals, quizzes, or mood trackers) and personal stories. Clients receive access to the iCBT program for up to 6 months; however, the core treatment is centred around 8 review periods (at intervals of 1 to 2 weeks), during which clinical supporters guide clients. These clinical supporters are a specially trained cohort of Psychological Wellbeing Practitioners (PWPs, https://www.instituteforapprenticeships.org/apprenticeship-standards/psychological-wellbeing-practitioner), typically graduate psychologists, with further training in low-intensity CBT-based interventions, including iCBT. Their support involves reviewing the engagement their client achieves from week to week and writing feedback to the client on their work, aiming to provide person-centred care that ensures good client experience and desired clinical outcomes.

All users in this study consented (oral or written) for their de-identified data to be used in analyses for routine service evaluation and improvement. Matching the terms and conditions of the service, and user consent, permits the analysis of anonymous data for research purposes and improves the treatment platform's effectiveness and service tools. Users were informed that this analysis might include profiling, machine learning or other techniques. Where individual-level data is used, additional safeguards such as anonymisation, use of pseudonyms (pseudonymisation) and limiting the set of individual data used (data minimization) are in place (SilverCloud Privacy notice: https://uk.silvercloudhealth.com/help/privacy/). Since this study does not obtain information through intervention or interaction with SilverCloud users and only uses secondary, fully de-identified data, it is not classified as human subjects research and, therefore, exempt from IRB review (Definition of Human Subjects Research: https://grants.nih.gov/policy/humansubjects/research.htm).

All steps taken to fully de-identify the data before any analysis are detailed in S1 File. This included the removal of free text entries, demographic information, and exact program dates. From this de-identified data, we selected clients who used the Space from Depression and Anxiety program, had an assigned clinical supporter, and completed clinical symptom measures at least twice. This resulted in 45,352 clients with regular reports of depression symptoms and 45,756 clients with regular reports of anxiety symptoms (Fig 1). In addition, to demonstrate the generalizability of our models, we also evaluated our models performance on 4 external datasets. These included one US (depression n = 2585, anxiety n = 2572) and two UK service providers (depression n = 41,774, anxiety = 41,667) as well as different iCBT programs (see S4 File for details). For one further, smaller UK single service program (n = 82 clients), we

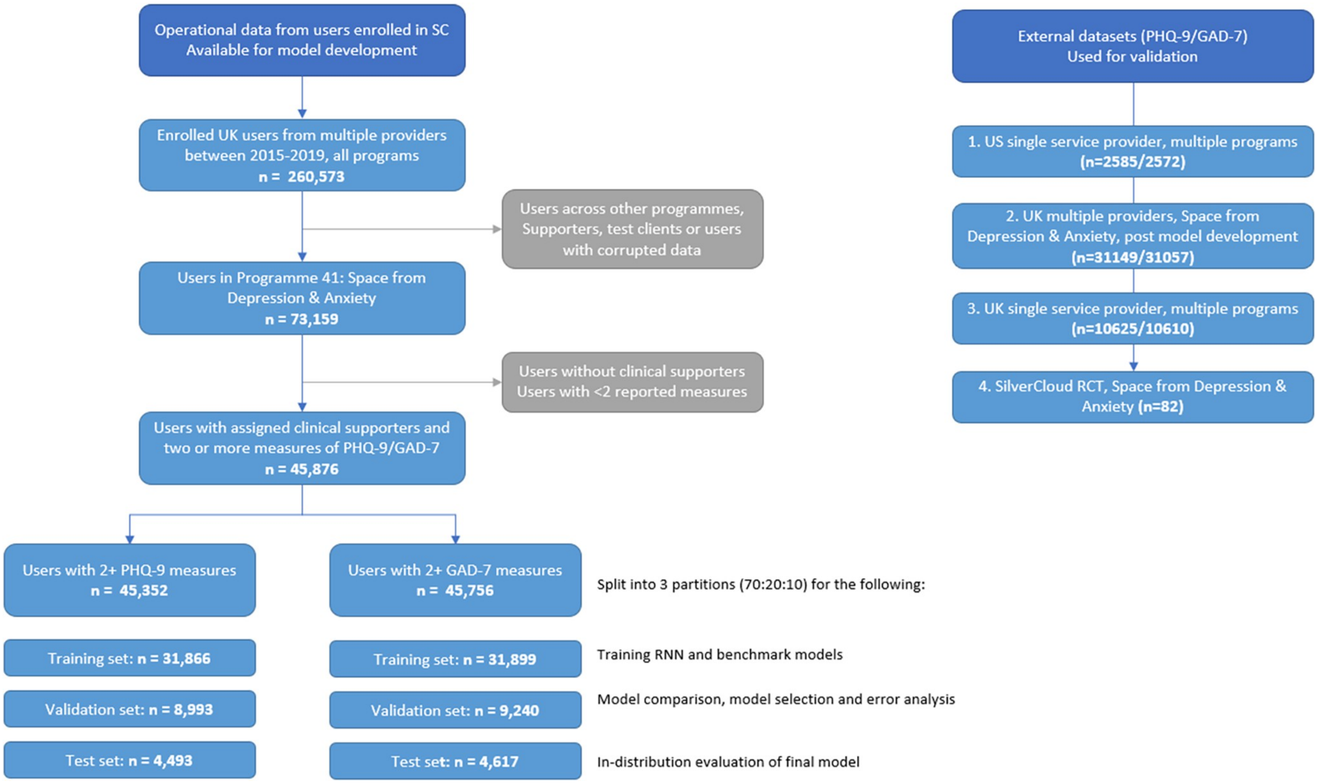

Operational data from users enrolled in SC
Available for model development
Enrolled UK users from multiple providers
between 2015-2019, all programs
n = 260,573
Users across other programmes,
Supporters, test clients or users
with corrupted data
Users in Programme 41: Space from
Depression & Anxiety
n = 73,159
Users without clinical supporters
Users with <2 reported measures
Users with assigned clinical supporters and
two or more measures of PHQ-9/GAD-7
n = 45,876
Users with 2+ PHQ-9 measures
n = 45,352
Users with 2+ GAD-7 measures
n = 45,756
Split into 3 partitions (70:20:10) for the following:
Training set: n = 31,866
Training set: n = 31,899
Training RNN and benchmark models
Validation set: n = 8,993
Validation set: n = 9,240
Model comparison, model selection and error analysis
Test set: n = 4,493
Test set: n = 4,617
In-distribution evaluation of final model
External datasets (PHQ-9/GAD-7)
Used for validation
1. US single service provider, multiple programs
(n=2585/2572)
2. UK multiple providers, Space from
Depression & Anxiety, post model development
(n=31149/31057)
3. UK single service provider, multiple programs
(n=10625/10610)
4. SilverCloud RCT, Space from Depression &
Anxiety (n=82)

Fig 1. Overview of all datasets used in the analysis. Left: CONSORT diagram of data inclusion pipeline; right: Summary of external datasets (additional detail can be found in S4 File).

https://doi.org/10.1371/journal.pone.0272685.g001

had demographic and clinical information for clients who participated in an RCT to test the effectiveness of SilverCloud Health. For all others, the nature of data access means that details on client diagnosis and other demographics such as their specific age, sex, race, ethnicity, or other socio-economic information were not available. Table 1 provides a structured overview of the key study elements following recent reporting guidelines for ML applications in clinical research [34, 35].

Pre-processing and feature engineering

For all clients included in the analysis, the available data include: (i) *timestamped scores from the completion of clinical measures* assessing symptoms of depression and anxiety, as well as (ii) *timestamped entries for each client interaction with the iCBT program*. Together, these form the input features (i.e., “independent variables”) that are fed to our DL framework for predicting client outcomes (see Table 2).

The clinical measures comprise both the *individual component questions* and *total scores* for the nine-item Patient Health Questionnaire (PHQ-9) for depression (0–27 total score) [36] and the seven-item Generalized Anxiety Disorder scale (GAD-7) for anxiety (0–21 total score) [37]. These are extensively validated self-report instruments employed in routine outcome monitoring (ROM) within IAPT services. Clients are asked to complete these instruments by the clinical supporter, usually every 1–2 weeks during treatment, and ask the client to indicate: “Over the last two weeks, how often have you been bothered by any of the following problems?”. The PHQ-9 problem statements include: “Feeling down, depressed or hopeless”; and for the GAD-7: “Feeling nervous, anxious or on edge”. Each question is responded to on an ordinal scale from 0 (*not at all)* to 3 (*nearly every day*), with higher scores indicating more severe depression or anxiety symptoms.

We had the following information for the client’s interactions with the iCBT program. We have data on the **section** of the program that the client used, which can be *Content* pages that the person views; the use of interactive therapy *Tools*; their viewing or writing of a *Journal*; visiting their user *Profile* page; or whether they interacted with their supporter as part of the *Review* process. In addition to the program section, the data logs also include the types of **actions** taken concerning these sections. For example, the client can ‘view’, ‘bookmark’ or ‘complete’ treatment *Content*. For ‘*Content’* and ‘*Tools’* actions, the data also included a **topic id** and **tool id specifying the content or tool** accessed. The combination of section-action-topic/tool id taken together (e.g. {*Tools*, ‘add’, id: Thought-Feelings-Behavior cycle}) define a set of 1133 possible unique interactions that a client can take in the iCBT program. Thus, as our final feature set, we have the *count of each* of the 1133 unique iCBT program interactions (i.e., how often the client reviewed content, engaged with therapeutic tools, and so forth) that a client performed within a given review period. We explain what constitutes a review period next.

Since interactions with the iCBT program accumulate the more the client engages with treatment, and clinical measures are frequently completed over time, we needed to identify meaningful time intervals to assess or predict the client’s progress. We chose *review periods* as our prediction points: the SilverCloud platform is configured such that a review period begins with the client completing one set of clinical measures (that is, PHQ-9/GAD-7) that are assigned to them by their clinical supporter. Review periods can therefore be seen as a proxy for time, on average 1.8 weeks in duration (see Fig 1 in S1 Appendix), though this can vary from client to client (SD = 0.24). At the end of the review period, the clinical supporter reviews the clinical measures and all other client interactions with the iCBT program. Based on their assessments, they provide personalised feedback to clients via an online message or telephone

Table 1. Overview of the key study elements.

Study design	
Clinical question	Can we predict early and robustly reliable improvement (RI) from iCBT treatment for depression and anxiety?
Model task/output	Binary classification of treatment outcomes (RI by end of program) given access to clinical measurements, iCBT interaction data, or both.
Intended use of results/Target user	Predictions produced by the model can be made available to clinical supporters as part of management, to assist treatment adaptation for improved patient outcomes
Study population + setting	
Population	Clients receiving iCBT treatment for symptoms of depression and anxiety
Study setting	NHS IAPT services in the UK
Data source	SilverCloud Health Platform
Cohort selection (Exclusion/ inclusion criteria)	Adult patients accessing the "Space from Depression and Anxiety Programme" with a clinical supporter. Client have at least 2 completed sets of PHQ-9 + GAD-7 measures within the enrolment period (96% of measures occur within the first 16 weeks)
Patient demographics	
Age	18+
Sex	Not provided
Race	Not provided
Ethnicity	Not provided
Socioeconomic status	Not provided
Data sources	
Data types	PHQ-9 and GAD-7 questionnaires; Interaction events with iCBT treatment
Data collection + transformation	See S1 File
Data structure + types	Ordinal (questionnaire scores); Count (interaction events)
Data partitions	See Fig 1
Model architecture	
ML methods & rationale	Recurrent neural networks (RNNs): can achieve high performances in modelling complex, high-dimensional data, and capturing long-term temporal dependencies in observations. Benchmarks: Logistic regression (LogR), Random forests (RF), Gradient boosting classifiers (GBMs), and Exponential moving averages (EMA)
Features	See Table 2
Data labels	Reliable improvement (see Table 3) between baseline (first) and last available (of a maximum of 8) self-reported measure for PHQ-9/GAD-7 respectively
Missingness	Last observation carried forward for label censoring due to user dropout before review 8
Hardware, software, packages	Model training + testing on the AzureML infrastructure, in Python 3.7 RNNs implemented via PyTorch
Data split	70:20:10 random split (training, validation and test)
Model training	RNN trained with cross-entropy loss for binary classification task. Hyperparameter tuning based on classification accuracy
Model parameters/ hyperparameters	3-layer RNN comprising 50-dim LSTM hidden layer + linear softmax. Dropout in LSTM layer with probability 0.4 Optimization using ADAM
Model evaluation/ validation	
Evaluation measures	Accuracy, AUROC, Sensitivity at fixed specificity
Internal model validation	Validation set (n = 9.2k) for model comparison and selection. Test set (n = 4.6k) for evaluation of final model
External model validation	Multiple external validation (see Table 6, S4 File)

(*Continued*)

Table 1. (Continued)

Transparency, reproducibility, code reuse	Data is not available. Code is available on request from the corresponding author

Key study elements listed according to recent reporting guidelines by Hernandez-Boussard et al. [34] and Stevens et al. [35] for applications of machine learning in clinical research.

https://doi.org/10.1371/journal.pone.0272685.t001

Table 2. Component features for predicting reliable improvement.

Feature Type	# Dimensions	Description
Previous Total PHQ-9	1	Scale 0–27; based on PHQ-9 scores between week 0 (baseline) and 8 weeks for each client
Previous Total GAD-7	1	Scale 0–21; based on GAD-7 scores between week 0 (baseline) and 8 weeks for each client
Component PHQ-9 questions	9	Scale 0–3
Component GAD-7 questions	7	Scale 0–3
iCBT Program Interactions	1133	Counts of each unique type of program interaction as defined by their *action*, *section*, and *topic/tool ID*

https://doi.org/10.1371/journal.pone.0272685.t002

contact. At the end of this process, clinical supporters assign questionnaires for a new set of clinical measures to the client, which marks the start of the following review period. We defined the end of treatment as a total of eight review periods into the program, which equates to an average of 13 weeks from the start of treatment (first review). We chose eight review periods as the upper limit since, although some users may engage with the platform beyond this time, they are no longer guaranteed a review by clinical supporters.

Taking the data described above as input features, our objective is the early prediction of treatment outcomes. Within NHS IAPT services in the UK, the "reliable improvement" outcome metric presents a core performance metric for determining treatment success. Reliable improvement reflects a significant positive change in client symptoms, based on the reliable change index by Jacobson and Truax [38]. It is defined as a decrease in total PHQ-9 score of 6 or more points, a decrease in total GAD-7 of 4 or more points, or both at the end of the user's program (Table 3) [39]. Reliable improvement is chosen as our outcome of interest as the definition indicates *real* improvement in symptoms, exceeding change that can be accounted for by measurement error. In addition, reliable improvement moves the assessment of clinical change from the group mean to the individual being treated.

For those users with measurements at fewer than eight review periods. we base the outcome on the last available measurement; we use last-observation-carried-forward (LCOF) imputation to handle label censoring due to user dropout. Evaluation of predictive performance is based only on time points up to and including the last available measurement for each user.

Table 3. Outcomes of interest definitions.

Outcome of Interest	Definition
Reliable improvement in depression	Decrease in PHQ-9 by ≥6 points with no increase ≥4 in GAD-7 at the end of the programme (8 review periods);
Reliable improvement in anxiety	Decrease in GAD-7 by ≥4 points with no increase ≥6 in PHQ-9 at the end of the programme (8 review periods)

https://doi.org/10.1371/journal.pone.0272685.t003

Overall, approximately 26% of users included in this study experience reliable improvement in PHQ-9 and 38% experience reliable improvement in GAD-7, where the majority of these are users with higher (moderate to severe) baseline symptom scores (see Fig 2 in S1 Appendix). As evidenced by our findings later, there is necessarily a trade-off between time to prediction and predictive accuracy, whereby earlier predictions may be most helpful to clinical decision-making yet more prone to model uncertainty and prediction errors.

Machine learning for outcome prediction

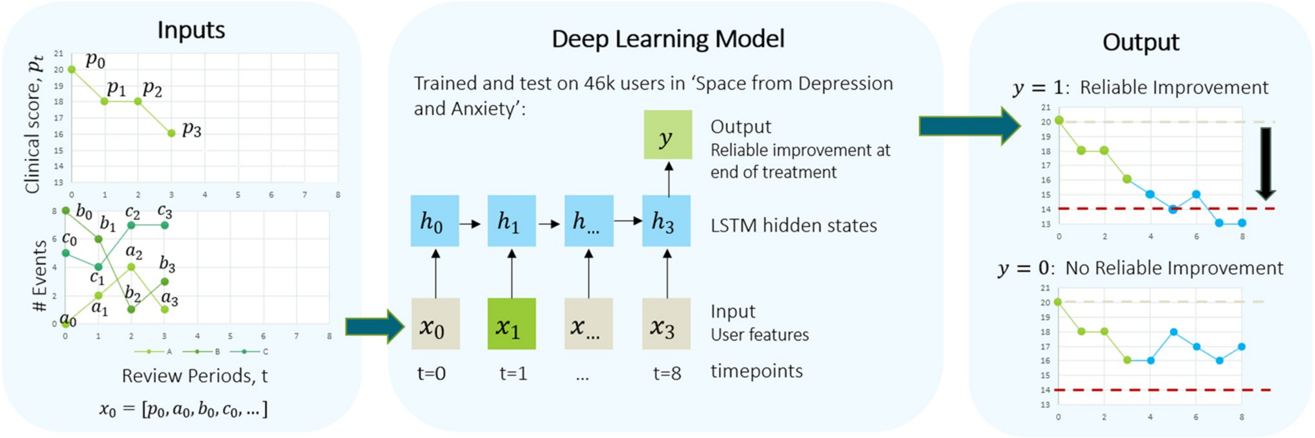
We use a deep learning framework to train and test the prediction of reliable improvement in depression and anxiety symptoms for a given client over the course of treatment. Specifically, we consider deep recurrent neural networks (RNNs). RNNs are a class of deep learning methods designed to model sequential data or time series data. Unlike traditional feedforward neural networks, RNNs have connections between the neurons that form directed cycles, allowing them to maintain a hidden state or "memory" of past inputs [40]. This makes them particularly well-suited to learning patterns and dependencies in sequences, and these models have been shown to achieve state-of-the-art performance in many applications with sequential data, including clinical time series [30, 41]. In contrast to classical linear models such as logistic regression, or iterative variants [20, 28], deep RNNs provide a natural framework for extracting patterns from high-dimensional, unstructured data, capturing long-term temporal dependencies in observations, and making predictions with variable length inputs [40]. This makes them well-suited to modelling rich, sequential clinical measurement and iCBT program interaction data. Additionally, the availability of a large-scale digital dataset makes it feasible to train and reliably evaluate these more data-intensive models.

We investigate RNN models trained on our outcomes of interest (reliable improvement in PHQ-9 and GAD-7, respectively), using four different combinations of the feature sets (Table 2):

1. Counts of *each* of the 1133 unique iCBT program interactions (**I**);
2. Early total PHQ-9 and/or GAD-7 questionnaire scores (**Q**);
3. Answers to individual PHQ-9 and GAD-7 questionnaire component question items (**Qc**); and
4. Combination of (i) iCBT program interactions and (ii) total clinical scores (**I+Q**).

Using these inputs, we split our data at random 70:20:10 into training, validation, and test sets, respectively, ensuring stratified samples by outcome label. We then trained a three-layer RNN to predict a binary outcome (whether a client will show reliable improvement at the end of the eight review periods (Fig 2)) using a many-to-one architecture, mapping a variable-length sequence of input data points to produce a single output or prediction. We used a 50-dimensional hidden layer with long-short term memory (LSTM) units to encode client features at each time point. Prediction models were trained independently for depression (PHQ-9) and anxiety (GAD-7) outcomes.

We also implemented several benchmark classification methods to compare against the predictive performance of the RNNs: logistic regression (LogR), random forests (RF), and gradient boosting classifiers (GBMs). These models form the basis of iCBT outcome prediction models in existing literature [24, 29, 42], and have been shown to achieve state-of-the-art performance. In each case, these models take as input feature vectors of length 16, comprising the total PHQ-9 or GAD-7 scores at all time points available so far (0 where not available to avoid imputation bias in the feature space and treating measurements as missing-at-random) up to a



a = Added Journal entry
b = Viewed topic "Boosting Behaviour"
c = Viewed tool "Activity Scheduling"

Fig 2. Deep learning architecture to predict reliable improvement in depression or anxiety. We used counts of engagement events with the SilverCloud platform and early clinical scores (PHQ-9 and GAD-7 for depression and anxiety, respectively) as the input features. Using these inputs, we trained a Recurrent Neural Network (RNN) to predict whether a client has reliable improvement at the end of the treatment period. We used LSTM hidden states to represent input features at each time point.

https://doi.org/10.1371/journal.pone.0272685.g002

maximum of 8 measures, and 8 indicator variables, denoting whether a measure is, in fact, available at that time. This allows us to make a fair comparison with RNNs trained on total scores alone (RNN-Q). Unlike RNNs, however, these models do not *explicitly* use the input data's sequential nature. As an additional benchmark, we implement exponential moving averages (EMA), a naïve trend-following algorithm, for predicting reliable improvement. The EMA forecasts the expected subsequent clinical measurement and uses this to generate a binary outcome label. The RNNs and benchmark models were built on the training set. Feature and model selection was based on the results from the validation set. The final model was evaluated on the held-out test set and several external cohorts. The models were trained and tested on the AzureML infrastructure in Python 3.7. RNNs were implemented using PyTorch.

Results

To investigate the most informative features in the prediction of treatment outcomes, we trained RNN models on the four different feature sets (I, Q, Qc, I+Q) described previously and reported the overall performance of these models on our validation dataset (Table 4). The RNN based on input features combining total clinical questionnaire scores (Q) and iCBT program interaction data (I) achieves the highest accuracy for both PHQ-9 (RNN-I+Q = 83.9%) and GAD-7 (RNN-I+Q = 79.84%), while interaction data alone (I) is the poorest predictor of reliable improvement (RNN-$I_{PHQ\text{-}9}$ = 71.15%, RNN-$I_{GAD\text{-}7}$ = 60.93% accuracy). Interestingly, the best-performing models (RNN-I+Q), built on significantly higher-dimensional data, only marginally outperform the RNNs built on total clinical scores alone (RNN-Q). Similarly, using individual score components (RNN-Qc) instead of the total score does not provide any gains to the prediction task, yielding slightly lower accuracies for PHQ-9 and GAD-7 outcome prediction.

To understand the clinical applicability of the predictive model, we also report performance at various operating points trading off sensitivity and specificity, focusing on regions of high specificity (minimizing false positives, where clients are wrongly predicted to improve). While

Table 4. Optimizing data featurization.

	PHQ-9 (n_{users} = 8993)						*GAD-7 (n_{users} = 9240)*					
	Acc	**Sens**	**Spec**	**90%**	**95%**	**97%**	**Acc**	**Sens**	**Spec**	**90%**	**95%**	**97%**
RNN-Q	83.75	49.85	97.55	61.91	53.08	50.66	78.86	61.1	91.08	61.1	54.2	47.62
RNN-Qc	83.66	50.25	97.26	62.44	54.28	50.65	78.8	62.68	89.89	62.53	55.17	48.0
RNN-I	71.15	1.63	99.46	18.98	10.44	6.51	60.93	20.67	88.65	18.33	10.1	6.53
RNN-I+Q	83.90	55.11	95.63	65.25	56.01	51.3	79.85	65.57	89.68	65.12	55.33	47.46

The table shows RNN validation set performance (in terms of accuracy (Acc.), sensitivity (Sens.), specificity (Spec.), along with sensitivity at fixed thresholds of 90%, 95% and 97% specificity respectively) with different feature inputs, averaged over all prediction timepoints. The combination of features considered are: overall PHQ and GAD scores (Q; input dimension = 1), individual score components (Qc; input dimension– 9/7), engagement alone (I; dimension = 1133) and engagement with total clinical scores (I+Q; dimensions = 1134).

https://doi.org/10.1371/journal.pone.0272685.t004

RNN-I+Q consistently performs best for PHQ-9 and GAD-7 improvement prediction, gains over RNN-Q are limited, particularly in regions of high specificity. This, together with the fact that RNN-Q is a much more parsimonious model with fewer data requirements, motivates the use of RNN-Q over RNN-I+Q in practice. We, therefore, restrict ourselves in subsequent analysis to models taking only the total questionnaire scores as input (RNN-Q).

To further compare the performance of the RNNs with three benchmark machine learning models: LogR, RF, GBMs, as well as the results of the EMA, Fig 3 plots the predictive accuracy of each of these models over time. We find that the RNN-Q consistently outperforms other ML benchmarks, with gains increasing with time into the program. While EMA—which explicitly models temporal trends—performs reasonably well (outperforming logistic regression and other machine learning methods), RNN-Q consistently achieves higher accuracies when three or more clinical measurements are available at the time of prediction.

Model generalization and subgroup performance

The overall predictive performance of RNN-Q on our test set for PHQ-9 (82.37%) and GAD-7 (77.92%) is summarized in Table 5; further details can be found in S2 File. Again, with the lens

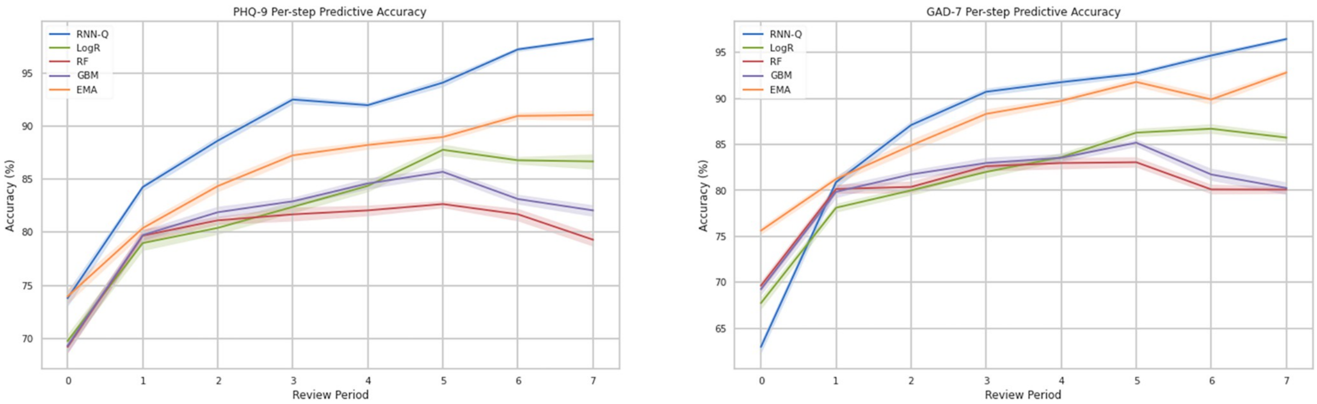


Fig 3. RNN-Q model performance on test dataset by review period plotted against benchmarks for both PHQ-9 and GAD-7. We evaluate our predictions from these two models by stratifying according to the number of measurements available at the time of prediction, to visualize their performance over time. Benchmarks of logistic regression (LogR), random forests (RF), gradient boosting machines (GBM) and exponential moving averages (EMA) are plotted alongside RNN-Q model performance for PHQ-9 (left) and GAD-7 (right), with bootstrapped confidence intervals. All of the above models have access to all prior clinical measures (up to the review period at prediction time). Original files can be found in S1 Dataset.

https://doi.org/10.1371/journal.pone.0272685.g003

Table 5. RNN-Q test performances for PHQ-9 and GAD-7 at different time intervals and across user subgroups.

		Accuracy		AUROC		Sens. at 95% Spec.	
		All t	**t≥3**	**All t**	**t≥3**	**All t**	**t≥3**
PHQ-9 (n_{users} = 4493)	**Overall**	**82.37**	**87.77**	**0.849**	**0.895**	**50.54**	**66.1**
	Mild	93.75	93.94	0.835	0.890	42.98	62.35
	Moderate	79.27	85.93	0.806	0.871	48.16	52.56
	Severe	75.01	84.26	0.814	0.877	47.7	51.62
GAD-7 (n_{users} = 4617)	**Overall**	**77.92**	**87.37**	**0.849**	**0.910**	**52.66**	**70.27**
	Mild	89.19	90.88	0.843	0.895	48.43	72.16
	Moderate	75.41	86.49	0.817	0.906	49.58	69.35
	Severe	73.69	86.23	0.826	0.895	50.5	54.47

RNN-Q test performance for accuracy, AUROC and sensitivity and 95% specificity, overall and by subgroups according to initial severity of symptoms. Performance is reported for both predictions at all time points t, and for only t ≥ 3 (when a minimum of three clinical measures are available for prediction).

https://doi.org/10.1371/journal.pone.0272685.t005

of clinical applicability, we investigate the model performance (i) across different client symptom subgroups; and (ii) at time points for which predictions will be made available. We found that prediction accuracy is higher for clients with mild depression and anxiety symptoms (with initial PHQ-9 between 0–8 or initial GAD-7 between 0–6) than those reporting moderate to severe baseline symptoms, which may be due to the class imbalance in this subgroup: few users with low initial scores will go on to meet the change criteria for reliable improvement. We also observed a largely monotonic improvement in the RNN's performance over time, with accuracy consistently above 87% across all client populations after three review periods. More detailed evaluation of model performance is presented in S3 File.

The solid predictive performance for both PHQ-9 and GAD-7 outcomes symptom subgroups motivates closer examination of the prediction errors that occur to interpret model behaviour. Figs 4 and 5 (for PHQ-9 and GAD-7, respectively) show the prediction errors for the models on the validation dataset, at review period t = 3 (i.e., given three measures so far in the program), for subgroups with different baseline symptom severity. We find that errors are generally concentrated in clients for whom the actual change in scores by the end of the program is near the threshold for reliable improvement, likely within the margin of measurement error; this is particularly pronounced in the mild/moderate groups. In addition, looking at the typical trajectory for false negatives in reliable improvement prediction for PHQ-9, we find that this typically occurs when there has been little improvement in scores until the point of prediction, followed by a sudden drop from the fourth measurement onwards. Conversely, false positives (which are in the minority of errors) occur in trajectories that see large improvements early in the program, but regress to much more modest improvement over the baseline as time progresses. In both cases, these are very much plausible errors—large drops or reversals in trends are naturally harder to predict—and provide further validation of model behaviour.

Finally, we evaluated the models performance on several external datasets across UK and US settings and for different iCBT programs. Table 6 provides an excerpt of the main findings; full results can be found in S4 File. These results are again compared against model performance on the internal test dataset and demonstrate the robustness of the achieved models across different health services, geographic locations, and iCBT programs. While demographic characteristics were largely unavailable, we also include model performance across age, gender and ethnicity subgroups on a small external cohort for which this data was collected and show that high accuracy is maintained across these subpopulations.

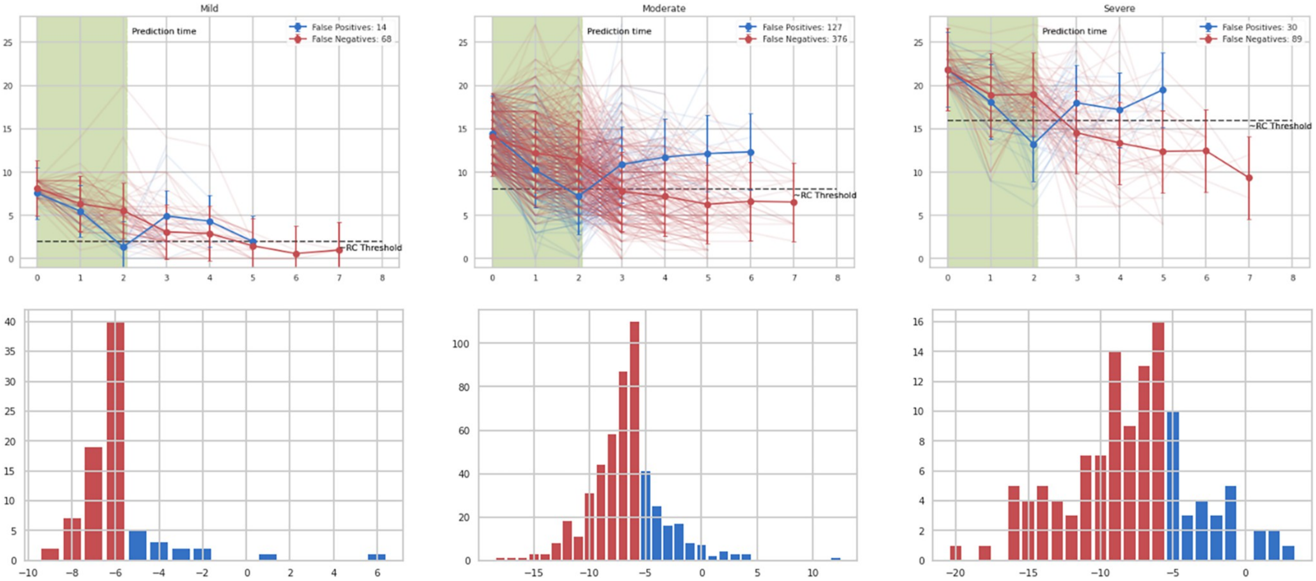


Fig 4. PHQ-9 error analysis. Top three figures (left-to-right) show the symptom trajectories for each of the errors, along with the mean trajectory, in bold. Bottom three figures (left-to-right) show the distribution of final changes in overall symptom score across the errors. Shaded in green are the measurements available during prediction (t = 3). False negatives (red) occur when symptom improvement is marginal in the initial weeks. False positives (blue) occur where a sharp fall in client-reported symptoms is observed before the prediction point, but scores rise again towards the baseline. In both modes of error, the true change in client scores is clustered around -6 (the threshold for reliable improvement), as would be expected, and tails off rapidly away from this threshold.

https://doi.org/10.1371/journal.pone.0272685.g004

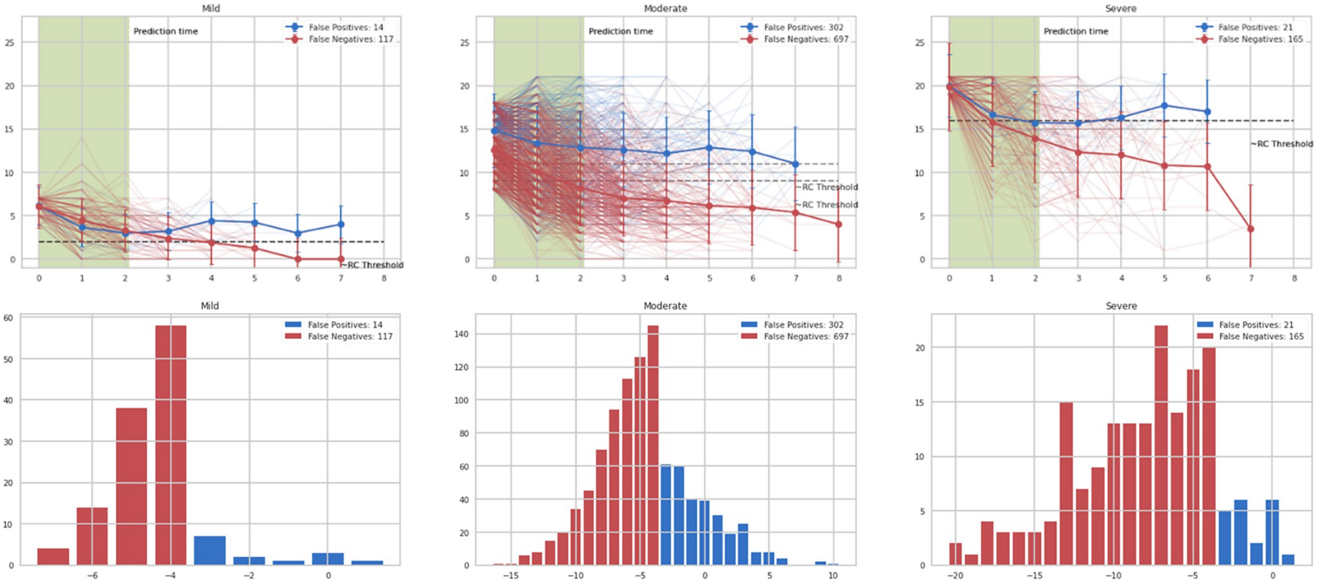


Fig 5. GAD-7 error analysis. Top three figures (left-to-right) show the symptom trajectories for each of the errors, along with the mean trajectory, in bold. Bottom three figures (left-to-right) show the distribution of final changes in overall symptom score across the errors. In both false positives (blue) and false negatives (red), the true change in client scores is clustered around -4 (the threshold for reliable improvement), as would be expected, and tails off rapidly away from this threshold.

https://doi.org/10.1371/journal.pone.0272685.g005

Table 6. Overview of model performance across external cohorts.

External Cohort Performance (all t)	*PHQ9*				*GAD7*			
	n_{users}	*Acc.*	*Sens.*	*Spec.*	n_{users}	*Acc.*	*Sens.*	*Spec.*
US single service provider *Multiple programs (D1)*	2585	86%	48%	97%	2572	78%	64%	87%
UK multiple providers; post development *"Space from Depression & Anxiety" (D2)*	31149	83%	50%	96%	31057	77%	65%	87%
UK single service provider *Multiple programs (D3)*	10625	82%	51%	96%	10610	77%	64%	87%

Model performance on external cohorts from varying time periods, geographic locations (US, UK), health service providers and iCBT programs.

https://doi.org/10.1371/journal.pone.0272685.t006

Discussion

The availability of coherent, high-fidelity digital engagement data (with click-level interactions logged for users all with access to the same treatment) and clinical outcome measures, as part of one of the largest datasets available for a clinical sample receiving internet-delivered cognitive behaviour therapy (iCBT), supported the development and evaluation of a deep learning framework for the prediction of treatment outcomes. This allowed us to counter many modelling limitations inherent to the classical statistics and machine learning approaches seen in prior work. Our best-performing models, based on total questionnaire scores of PHQ-9 and GAD-7, consistently predicted reliable improvement changes in clinical symptoms with above 87% accuracy and 0.89 AUC early during treatment (after three or more review periods). Notably, we achieve above 66% sensitivity for PHQ-9 and 70% for GAD-7 when designing for over 95% specificity (a false positive rate of 1 in 20), which is crucial to the utility of these predictions in clinical decision-making. While the limited prior work with large-scale observational iCBT data and variations in client selection criteria or outcome of interest makes direct comparisons challenging (most recently, Bone et al. achieved an AUC of 0.81 at a comparable stage in treatment) [24], to our knowledge this is the best performance achieved in the dynamic prediction of psychological treatment outcomes to date–especially for predicting clients who are at risk of not achieving RI [17]–and the first use of deep learning for outcome prediction from a dataset treating clients using iCBT alone (for other iCBT-based explorations in this space see Boman et al. [43]). Our RNN-based approach outperformed several competitive benchmarks and maintained predictive accuracy across multiple large-scale external cohorts, supporting its utility in outcome prediction across diverse populations, different iCBT programs, and geographical locations (UK, US). This latter work increases the ecological validity of the DL model, work that has not been achieved by previous developments in outcome prediction in this area [17].

Having explored different feature sets in training RNN models, we found that the RNN based on input features combining total clinical questionnaire scores (Q) and iCBT program interaction data (I) achieved highest accuracy, outperforming those built on either one dimension or using individual score components instead of total scores. This may be attributed in part to increased overfitting, as well as the significant heterogeneity across the client population in symptom profiles and, in turn, which symptom changes contribute most to reliable improvement for different client subgroups, making it challenging to outperform the use of aggregate scores in predicting the defined outcomes. We also found in our analysis that iCBT program interactions, though a rich source of data on the platform, yield noisy predictors when taken alone as input, and contribute little additional predictive value to use of just clinical measures. One reason may be that program interactions, as a proxy for treatment engagement, contribute to outcomes via psychological change. This contribution itself will be represented within the outcome data of previous weeks, limiting the value of including interaction data separately. There may also be opportunities for different approaches to learning representations for program interactions. In our work, we encoded these as 1133 unique section-action-content sequences, which may not be the right granularity to extract meaningful patterns. Possible directions for future work include defining interactions based solely on the treatment section, grouping therapy contents by overall theme, or identifying the unique treatment components that contribute to behaviour change.

On the other hand, information may be lost by representing program interaction solely on aggregate counts: modelling instead of temporal patterns in platform use within each review period may be more informative; the RNN architecture provides a flexible, high-capacity framework for such exploration. Furthermore, we must acknowledge that within this digital treatment set-up, it can be harder to separate between a client's interactions with the digital platform and their engagement with treatment (i.e., client 'views' of therapy-based *Content* pages themselves carry little information about the extent to which the client may attentively read and understand that content). While the relationship between platform interactions and outcomes is likely complex [44], active engagement with the iCBT program is essential in contributing to change. It is well understood and often targeted by clinicians to help improve client outcomes [45]. Thus, further research into including treatment interaction for explainable predictions is warranted.

Results illustrate how the accuracy of the dynamic outcome predictions with our RNN models increases with time. While early prediction of client outcomes may be most informative to clinical practice, later predictions are more reliable. Given this trade-off, we suggest that predictions could be made available to PWPs after a minimum of three review periods (typically within the first six weeks), from which point our models consistently achieve above 87% accuracy and 0.89 AUROC whilst leaving a clinically actionable time window for therapy that typically lasts eight review periods. However, at the same time, this restriction to a minimum of three review periods means that a significant proportion of clients (almost 60% based on training data distribution), who might drop out, change or complete the program ahead of this time, would not be predicted for in practice. Future studies are needed to explore appropriate "prediction accuracy–client inclusion" trade-offs, and how they impact clinical utility during the provision of support, especially in the broader context of feedback-informed psychotherapy, where expected treatment response models typically require just two assessments [14].

In any clinical application of machine learning, we need to be mindful of the potential harm of prediction errors. In the context of feedback provision within iCBT, there is general agreement among clinical experts that false positives must be minimized, as these constitute clients that need extra help and may be at risk of not receiving it, potentially delaying recovery. On the other hand, false negative predictions mean that the model predicts that a client not to achieve reliable improvement when in fact, they do. False negatives can disrupt a client's treatment journey if they cause unnecessary or unhelpful adaptations–for example, the client is referred to more intensive care when they didn't need to be–leading to poorer client experiences and misdirection of limited resources. However, for most negative RI predictions (whether true or false), clinicians described that they would work harder to identify the clients' difficulties and treatment needs, meaning that false negative errors may not come to weigh in as much as false positives (not receiving extra help when needed), which is why we chose to restrict our model specificity to above 95% (low false positive rate), at the expense of a higher false negative rate. Despite such adjustments and overall low error rates, it remains paramount that clinicians are carefully educated about the probabilistic nature of prediction models and their potential for errors so that they can appropriately interpret and use the provided information. This is particularly important given the heterogeneity in treatment response and potential fluctuations at an individual level, particularly in cases where abrupt changes in symptom trajectories may occur, requiring clinicians to balance assessments of the predicted outcomes with their professional expertise. Future research must carefully assess the appropriateness and real-world implications of this chosen threshold/ trade-off and guide necessary adaptations.

Limitations

There are multiple ways in which outcomes can be defined to inform clinical decision-making including risks of symptom deterioration; chance of recovery or remission; specific score change; mental health trend; or, as in this case, a significant reduction in symptoms. The focus on reliable improvement was a deliberate choice as it is an established metric for measuring the success of treatments delivered in the IAPT program, the context in which this work is focused. However, we acknowledge that there are controversies and a lack of agreement in determining a good threshold for assessing improvement in mental health. Further, we decided to predict a score change ‘threshold’ rather than the change in the score itself (a more sensitive measure). We found the threshold to provide a less ambiguous signal that a client might be at risk of not improving significantly (negative RI prediction). Furthermore, in a subsequent, planned deployment study (see below), we decided not to present predictions for patients with scores below ‘caseness’, for whom it is more difficult (numerically) to achieve RI, which makes negative RI predictions most probable. Yet, those predictions are less likely to indicate poor patient progress or need for more support, as these individuals are already in the desired score bounds of recovery or remission.

Furthermore, with the prediction models based solely on clinical scores, we acknowledge that no further insights are given into the potential mechanisms of suggested treatment failures, nor are concrete actions proposed for how clinical supporters may address any treatment difficulties. Nonetheless, and in line with other recent research on the dynamic prediction of treatment outcomes [24], we consider a DL tool with high accuracy for predicting outcomes that serve as a ‘prompt’ for clinical review to significantly enhance response rates to treatment by enabling accurate and timely feedback to clinical supporters to improve clinical decision making (i.e., considering stepping clients-up or -out to alternative treatments, and tailoring treatment to client preferences and expectations) [17]. As such, this work responds to calls by intervention researchers to systematically evaluate client response to treatment to determine if the course is progressing as expected and, if not, to modify or change treatments as suitable [46].

Another limitation inherent to the observational setting of this study is that like previous works [17], our analyses rely on the use of last observation carried forward methods, where the last observed measure for each user is taken to determine the post-treatment outcome, which introduces label error for those clients who drop out before the end of the treatment, potentially conflating predictive performance. In S3 File, we quantify this gap by evaluating performance on only those users for which eight or more clinical measurements are available. However, given that our objective here is to evaluate the utility of dynamic predictions in clinical decision-making for *all* users with highly heterogenous platform use (many of whom may complete treatment content or achieve reliable improvement before the full intended length of treatment), our priority was to be as inclusive as possible in our analysis, requiring a minimum of just two clinical measurements. Additionally, there is considerable scope for further work to better understand outcome prediction over the course of treatment with this richer dataset. For example, previous psychotherapy research has shown how clients who improve early in therapy and make sudden mental health gains [47] tend to show the best outcomes. These effects are often demonstrated in conventional psychotherapy and have also been found for briefer, low-intensity CBT treatments [17]. If early change is a significant predictor of outcomes, future work could assess if there are any differences in model performance when predicting early versus later improvers, suggesting higher accuracy rates for those early improvers.

Towards real-world implementation

In addition to considerations of model robustness, several practical implications need to be realized to introduce machine learning insights successfully into real-world clinical workflows [48]. We regard iCBT as particularly well-suited to the adoption of these methods. In these digital health services, outcome data is already routinely collected and monitored, and clinical supporters are responsible for examining this outcome data in the context of iCBT treatment by using data dashboards to review client progress and support case management. It is thus more straightforward to present relevant insights to clinicians, focusing on the added value of these techniques and appropriate practices for their use, rather than implementation barriers. In a separate paper [49], we report initial learnings from user research with iCBT supporters that clarified concrete use scenarios and potential concerns about integrating achieved prediction outcome models within iCBT practice. It also highlights how design choices in the supporter user interface and workflow integration can help mitigate the risks of over-reliance on AI outputs.

The integration of the prediction models within the supporter user interface further paves the way for real-world studies to assess the impact of introducing robust and reliable DL methods into clinical practice, intending to support clinicians in making early and necessary treatment adjustments to promote a favourable therapeutic response in the greatest number of clients. Following on from this research, a randomized controlled trial (https://www.isrctn.com/ISRCTN18059067) is being designed to investigate the effectiveness of feedback-informed iCBT treatment through DL algorithms in improving symptoms of depression and anxiety, delving into clinicians' perception on acceptance of these models, their subjective experience with the model outputs, and ultimately their potential to impact the provision of support and improve clinical outcomes.

Conclusions

Multiple studies assessing the benefits of feedback-informed therapy (FIT) have demonstrated how providing therapists or clinical supporters who assist the person undergoing treatment with access to feedback on expected client outcomes from treatment can improve treatment success and prevent deterioration, especially in clients with poor prognoses. Expected treatment response models are the methodological standard for assessing treatment outcomes in psychotherapy interventions, with few studies investigating the use of machine learning for the dynamic prediction of desired treatment outcomes. Within iCBT contexts, predictive models have been built on modest and selective samples (ranging from <100 to a few thousand users) and have typically involved post hoc analysis of RCT data and miscellaneous data sources, with mixed results. Reported accuracies range between 55–83%; however, different means of computing those and variations in the underlying treatment programs make any direct comparisons challenging.

This study demonstrates the feasibility of using a deep learning (DL) framework to achieve robust, dynamic outcome prediction models. DL methods achieve state-of-the-art performance in many settings, given their ability to handle high-dimensional inputs and model complex, non-linear patterns in unstructured data with limited feature engineering. Deep RNNs provide a natural framework for modelling and making predictions given sequential data. The models presented here, based solely on routinely collected clinical outcome measures of depression and anxiety symptoms, yield high accuracies early into treatment (>87% after just three review periods), outperforming several other machine learning and advanced statistical methods. They also show generalizability to data from different iCBT programs, geographies, and demographic groups.

Methodological implications for the current work depart from previous limitations in evidence in two ways: first, by using a large-scale dataset in training and validation that closely matches the intended implementation context and second, by employing a high-capacity DL framework to deliver robust, dynamic outcome prediction with high accuracy. The clinical implications of these models in digitally delivered psychotherapy services include enabling the near-term deployment and clinical study of such models within iCBT care that already routinely collect clinical outcome measures and employ digital review practices. This moves the field of precision psychiatry one step closer alongside many paths that seek to enable clinical supporters to prioritize better and adapt real-time interventions for clients with the greatest unmet need.

Supporting information

S1 File. De-identification and pre-processing pipeline.
(DOCX)

S2 File. Model featurization in benchmarks vs RNN.
(DOCX)

S3 File. Internal validation and test performance results.
(DOCX)

S4 File. External validation results.
(DOCX)

S1 Appendix. Fig 1. User distribution and reliable improvement over time. Top figure: Distribution of time from enrolment to nth review; bottom figure: Counts of users at each review period. **Fig 2**. Kaplan Meier curves modelling reliable improvement events in PHQ-9 and GAD-7 for clients with different baseline severity. Survival probability corresponds to the probability that sustained reliable improvement is not achieved by a given timestep in the treatment program. A steeper drop in survival probability can be interpreted as higher rate of reliable improvement in that subpopulation.
(DOCX)

S1 Dataset. Benchmark numerical results on validation dataset for PHQ and GAD. These data sets correspond to the results shown in Fig 3 (per-step predictive accuracy).
(ZIP)

Acknowledgments

We thank our broader team: James Bligh and Catalina Cumpanasoiu.

Author Contributions

Conceptualization: Angel Enrique, Jorge Palacios, Derek Richards, Gavin Doherty, Danielle Belgrave, Anja Thieme.

Data curation: Tim Regan.

Formal analysis: Niranjani Prasad, Isabel Chien, Ryutaro Tanno, Hannah Richardson, Aditya Nori.

Methodology: Niranjani Prasad, Tim Regan, Usman Munir, Ryutaro Tanno, Danielle Belgrave, Anja Thieme.

Software: Dessie Keegan.

Supervision: Angel Enrique, Jorge Palacios, Derek Richards, Gavin Doherty.

Validation: Tim Regan, Dessie Keegan, Anja Thieme.

Visualization: Anja Thieme.

Writing – original draft: Niranjani Prasad, Isabel Chien, Gavin Doherty, Danielle Belgrave, Anja Thieme.

Writing – review & editing: Angel Enrique, Jorge Palacios, Derek Richards.

References

1. World Health Organization. Depression and other common mental disorders: global health estimates. World Health Organization; 2017.
2. O'Donohue WT, Fisher JE, editors. Cognitive behavior therapy: Core principles for practice. John Wiley & Sons; 2012 Jun 13.
3. Wright JH, Owen JJ, Richards D, Eells TD, Richardson T, Brown GK, et al. Computer-assisted cognitive-behavior therapy for depression: a systematic review and meta-analysis. The Journal of clinical psychiatry. 2019 Mar 19; 80(2):0-. https://doi.org/10.4088/JCP.18r12188 PMID: 30900849
4. Andrews G, Basu A, Cuijpers P, Craske MG, McEvoy P, English CL, et al. Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: an updated meta-analysis. Journal of anxiety disorders. 2018 Apr 1; 55:70–8. https://doi.org/10.1016/j.janxdis.2018.01.001 PMID: 29422409
5. Eilert N, Enrique A, Wogan R, Mooney O, Timulak L, Richards D. The effectiveness of Internet-delivered treatment for generalized anxiety disorder: An updated systematic review and meta-analysis. Depression and Anxiety. 2021 Feb; 38(2):196–219. https://doi.org/10.1002/da.23115 PMID: 33225589
6. Karapanos E. Sustaining user engagement with behavior-change tools. Interactions. 2015 Jun 25; 22 (4):48–52.
7. Karyotaki E, Efthimiou O, Miguel C, Bermpohl FM, Furukawa TA, Cuijpers P, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA psychiatry. 2021 Apr 1; 78(4):361–71. https://doi.org/10.1001/jamapsychiatry.2020.4364 PMID: 33471111
8. Schueller SM, Tomasino KN, Mohr DC. Integrating human support into behavioral intervention technologies: The efficiency model of support. Clinical Psychology: Science and Practice. 2017 Mar; 24(1):27.
9. Hayes SC, & Hofman SG (Eds.). 2018. Process-based CBT: The science and core clinical competencies of cognitive behavior therapy. Child & Family Behavior Therapy. New Harbinger Publications, Inc.
10. Hundt NE, Mignogna J, Underhill C, Cully JA. 2013. The relationship between use of CBT skills and depression treatment outcome: A theoretical and methodological review of the literature. Behavior Therapy, 44(1), 12–26. https://doi.org/10.1016/j.beth.2012.10.001 PMID: 23312423
11. Prescott DS, Maeschalck CL, Miller SD, editors. Feedback-informed treatment in clinical practice: Reaching for excellence. Washington, DC: American Psychological Association; 2017.
12. Lambert M. J., Hansen N. B., & Finch A. E. (2001). Patient-focused research: Using patient outcome data to enhance treatment effects. Journal of consulting and clinical psychology, 69(2), 159. PMID: 11393594
13. Amble I, Gude T, Stubdal S, Andersen BJ, Wampold BE. 2015. The effect of implementing the Outcome Questionnaire-45.2 feedback system in Norway: A multisite randomized clinical trial in a naturalistic setting. Psychotherapy Research, 25(6), 669–677. https://doi.org/10.1080/10503307.2014.928756 PMID: 25101527

14. Delgadillo J, Overend K, Lucock M, Groom M, Kirby N, McMillan D, et al. 2017. Improving the efficiency of psychological treatment using outcome feedback technology. Behaviour Research and Therapy, 99, 89–97. https://doi.org/10.1016/j.brat.2017.09.011 PMID: 29024821

15. Delgadillo J, de Jong K, Lucock M, Lutz W, Rubel J, Gilbody S, et al. Feedback-informed treatment versus usual psychological treatment for depression and anxiety: a multisite, open-label, cluster randomised controlled trial. The Lancet Psychiatry. 2018 Jul 1; 5(7):564–72. https://doi.org/10.1016/S2215-0366(18)30162-7 PMID: 29937396

16. Shimokawa K., Lambert M. J., & Smart D. W. (2010). Enhancing treatment outcome of patients at risk of treatment failure: Meta-analytic and mega-analytic review of a psychotherapy quality assurance system. Journal of Consulting and Clinical Psychology, 78, 298–311. https://doi.org/10.1037/a0019247 PMID: 20515206

17. Delgadillo J, McMillan D, Lucock M, Leach C, Ali S, Gilbody S. Early changes, attrition, and dose-response in low intensity psychological interventions. Br J Clin Psychol. 2014; 53(1):114–130. https://doi.org/10.1111/bjc.12031 PMID: 24117962

18. Muir HJ, Coyne AE, Morrison NR, Boswell JF, Constantino MJ. Ethical implications of routine outcomes monitoring for patients, psychotherapists, and mental health care systems. Psychotherapy. 2019; 56 (4):459–469. https://doi.org/10.1037/pst0000246 PMID: 31580139

19. Beshai S, Dobson KS, Bockting CL, Quigley L. Relapse and recurrence prevention in depression: current research and future prospects. Clinical Psychology Review. 2011 Dec 1; 31(8):1349–60. https://doi.org/10.1016/j.cpr.2011.09.003 PMID: 22020371

20. Ali S, Rhodes L, Moreea O, McMillan D, Gilbody S, Leach C, et al. How durable is the effect of low intensity CBT for depression and anxiety? Remission and relapse in a longitudinal cohort study. Behaviour research and therapy. 2017 Jul 1; 94:1–8. https://doi.org/10.1016/j.brat.2017.04.006 PMID: 28437680

21. Forsell E, Isacsson N, Blom K, Jernelöv S, Ben Abdesslem F, Lindefors N, et al. Predicting treatment failure in regular care Internet-Delivered Cognitive Behavior Therapy for depression and anxiety using only weekly symptom measures. Journal of consulting and clinical psychology. 2020 Apr; 88(4):311. https://doi.org/10.1037/ccp0000462 PMID: 31829635

22. Forsell E, Jernelöv S, Blom K, Kraepelien M, Svanborg C, Andersson G, et al. Proof of concept for an adaptive treatment strategy to prevent failures in internet-delivered CBT: a single-blind randomized clinical trial with insomnia patients. American Journal of Psychiatry. 2019 Apr 1; 176(4):315–23. https://doi.org/10.1176/appi.ajp.2018.18060699 PMID: 30696270

23. Lorenzo-Luaces L, DeRubeis RJ, van Straten A, Tiemens B. A prognostic index (PI) as a moderator of outcomes in the treatment of depression: A proof of concept combining multiple variables to inform risk-stratified stepped care models. Journal of Affective Disorders. 2017 Apr 15; 213:78–85. https://doi.org/10.1016/j.jad.2017.02.010 PMID: 28199892

24. Bone C, Simmonds-Buckley M, Thwaites R, Sandford D, Merzhvynska M, Rubel J, et al. Dynamic prediction of psychological treatment outcomes: development and validation of a prediction model using routinely collected symptom data. The Lancet Digital Health. 2021 Apr 1; 3(4):e231–40. https://doi.org/10.1016/S2589-7500(21)00018-2 PMID: 33766287

25. Chekroud AM, Bondar J, Delgadillo J, Doherty G, Wasil A, Fokkema M, et al. The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry. 2021 Jun; 20(2):154–70. https://doi.org/10.1002/wps.20882 PMID: 34002503

26. Hilbert K, Kunas SL, Lueken U, Kathmann N, Fydrich T, Fehm L. Predicting cognitive behavioral therapy outcome in the outpatient sector based on clinical routine data: a machine learning approach. Behaviour research and therapy. 2020 Jan 1; 124:103530. https://doi.org/10.1016/j.brat.2019.103530 PMID: 31862473

27. Flygare O, Enander J, Andersson E, Ljótsson B, Ivanov VZ, Mataix-Cols D, et al. Predictors of remission from body dysmorphic disorder after internet-delivered cognitive behavior therapy: a machine learning approach. BMC psychiatry. 2020 Dec; 20:1–9.

28. Lenhard F, Sauer S, Andersson E, Månsson KN, Mataix-Cols D, Rück C, et al. Prediction of outcome in internet-delivered cognitive behaviour therapy for paediatric obsessive-compulsive disorder: A machine learning approach. International Journal of Methods in Psychiatric Research. 2018 Mar; 27(1):e1576. https://doi.org/10.1002/mpr.1576 PMID: 28752937

29. van Breda W, Bremer V, Becker D, Hoogendoorn M, Funk B, Ruwaard J, et al. Predicting therapy success for treatment as usual and blended treatment in the domain of depression. Internet interventions. 2018 Jun 1; 12:100–4. https://doi.org/10.1016/j.invent.2017.08.003 PMID: 29862165

30. Rajkomar A, Oren E, Chen K, Dai A, Hajaj N, Liu P, et al. Scalable and accurate deep learning with electronic health records. npj Digital Med 1, 18 (2018). https://doi.org/10.1038/s41746-018-0029-1 PMID: 31304302

31. UAE. 2020. About IAPT and the History of the Programme: Stepped Care Model Information. Last retrieved 22nd June 2020 https://www.uea.ac.uk/medicine/departments/psychological-sciences/cognitive-behavioural-therapy-training/-about-iapt-and-the-history-of-the-programme/stepped-care-model-information

32. Mental health support for nurses: free access to online support (2021) Last accessed 23 December 2021 https://rcni.com/nursing-standard/newsroom/news/mental-health-support-nurses-free-access-to-online-support-and-no-referral-needed-173521

33. Online Behavioral Health Solutions for NHS | SilverCloud Health (2021) Last accessed 22 December 2021 https://www.silvercloudhealth.com/uk/online-behavioral-health-solutions-for-nhs

34. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc. 2020; 27(12):2011–2015. https://doi.org/10.1093/jamia/ocaa088 PMID: 32594179

35. Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for Reporting Machine Learning Analyses in Clinical Research. Circ Cardiovasc Qual Outcomes. 2020; 13(10):e006556. https://doi.org/10.1161/CIRCOUTCOMES.120.006556 PMID: 33079589

36. Kroenke Kurt, Spitzer Robert L., and Williams Janet BW. 2001. The PHQ-9: validity of a brief depression severity measure. Journal of general internal medicine 16(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x PMID: 11556941

37. Löwe Bernd, Decker Oliver, Müller Stefanie, Brähler Elmar, Schellberg Dieter, Herzog Wolfgang, and Herzberg Philipp Yorck. 2008. Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. Medical care, 266–274. https://doi.org/10.1097/MLR.0b013e318160d093 PMID: 18388841

38. Jacobson Neil S., and Truax Paula. "Clinical significance: a statistical approach to defining meaningful change in psychotherapy research." (1992).

39. Clark DM. Improving Access to Psychological Therapies Manual (updated). London: NHS England. 2021.

40. Lipton, ZC, Berkowitz, J, Elkan C.2015. A Critical Review of Recurrent Neural Networks for Sequence Learning. arXiv preprint arXiv:1506.00019.

41. Lipton ZC, Kale DC, Elkan C, Wetzel R. Learning to diagnose with LSTM recurrent neural networks. arXiv 2015. arXiv preprint arXiv:1511.03677.

42. Wallert J, Boberg J, Kaldo V, Mataix-Cols D, Flygare O, Crowley JJ, et al. Predicting remission after internet-delivered psychotherapy in patients with depression using machine learning and multi-modal data. Translational Psychiatry. 2022; 12, 357. https://doi.org/10.1038/s41398-022-02133-3 PMID: 36050305

43. Boman M, Ben Abdesslem F, Forsell E, Gillblad D, Görnerup O, Isacsson N, et al. Learning machines in Internet-delivered psychological treatment. Progress in Artificial Intelligence. 2019 Dec; 8(4):475–85.

44. Chien I, Enrique A, Palacios J, Regan T, Keegan D, Carter D, et al. A machine learning approach to understanding patterns of engagement with internet-delivered mental health interventions. JAMA network open. 2020 Jul 1; 3(7):e2010791-. https://doi.org/10.1001/jamanetworkopen.2020.10791 PMID: 32678450

45. Enrique A, Palacios JE, Ryan H, Richards D. Exploring the relationship between usage and outcomes of an internet-based intervention for individuals with depressive symptoms: secondary analysis of data from a randomized controlled trial. Journal of medical Internet research. 2019; 21(8):e12775. https://doi.org/10.2196/12775 PMID: 31373272

46. Duncan Barry L. "The Partners for Change Outcome Management System (PCOMS): The Heart and Soul of Change Project." *Canadian Psychology/Psychologie canadienne* 53, no. 2 (2012): 93.

47. Tang T. Z., & DeRubeis R. J. (1999). Sudden gains and critical sessions in cognitive-behavioral therapy for depression. Journal of Consulting and Clinical Psychology, 67, 894–904. https://doi.org/10.1037//0022-006x.67.6.894 PMID: 10596511

48. Thieme A, Belgrave D, Doherty G. Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. ACM Transactions on Computer-Human Interaction (TOCHI). 2020 Aug 17; 27(5):1–53.

49. Thieme A, Hanratty M, Lyons M, Palacios JE, Marques R, Morrison C, et al. 2023. Designing Human-Centered AI for Mental Health: Developing Clinically Relevant Applications for Online CBT Treatment. ACM Transactions on Computer-Human Interaction (TOCHI). 2023 30; 2(27):1–50.