Psychology

Self-study CPD activity

Activity Reference Number
C8(25) 2028

Topic

Deep learning model

Article(s)

Incorporating a deep-learning client outcome prediction tool as feedback in supported internet-delivered cognitive behavioural therapy for depression and anxiety: A randomised controlled trial within routine clinical practice

Profession(s)

All professions

This activity is approved for **THREE (3)** ***Clinical*** Continuing Education Units (CEU's)

Received: 30 November 2023 | Revised: 1 May 2024 | Accepted: 2 May 2024

DOI: 10.1002/capr.12771

ORIGINAL ARTICLE

WILEY

Incorporating a deep-learning client outcome prediction tool as feedback in supported internet-delivered cognitive behavioural therapy for depression and anxiety: A randomised controlled trial within routine clinical practice

Garrett C. Hisler | Katherine S. Young | Diana Catalina Cumpanasoiu | Jorge E. Palacios | Daniel Duffy | Angel Enrique | Dessie Keegan | Derek Richards

Amwell Science, Amwell, Boston, MA, USA

Correspondence
Garrett C. Hisler, Amwell Science, Amwell, 75 State St 26th floor, Boston, MA 02109.
Email: garrett.hisler@amwell.com

Abstract

Introduction: Machine learning techniques have been leveraged to predict client psychological treatment outcomes. Few studies, however, have tested whether providing such model predictions as feedback to therapists improves client outcomes. This randomised controlled trial examined (1) the effects of implementing therapist feedback via a deep-learning model (DLM) tool that predicts client treatment response (i.e., reliable improvement on the Patient Health Questionnaire-9 [PHQ-9] or Generalized Anxiety Disorder-7 [GAD-7]) to internet-delivered cognitive behavioural therapy (iCBT) in routine clinical care and (2) therapist acceptability of this prediction tool.

Methods: Fifty-one therapists were randomly assigned to access the DLM tool (vs. treatment as usual [TAU]) and oversaw the care of 2394 clients who completed repeated PHQ-9 and GAD-7 assessments.

Results: Multilevel growth curve models revealed no overall differences between the DLM tool vs. TAU conditions in client clinical outcomes. However, clients of therapists with the DLM tool used more tools, completed more activities and visited more platform pages. In subgroup analyses, clients predicted to be 'not-on-track' were statistically significantly more likely to have reliable improvement on the PHQ-9 in the DLM vs. TAU group. Therapists with access to the DLM tool reported that it was acceptable for use, they had positive attitudes towards it, and reported it prompted greater examination and discussion of clients, particularly those predicted not to improve.

Conclusion: Altogether, the DLM tool was acceptable for therapists, and clients engaged more with the platform, with clinical benefits specific to reliable improvement on the PHQ-9 for not-on-track clients. Future applications and considerations for implementing machine learning predictions as feedback tools within iCBT are discussed.

Garrett C. Hisler, Katherine S. Young shared authorship.

KEYWORDS
deep learning, feedback-informed treatment, machine learning, not-on-track, SilverCloud, internet-delivered cognitive behavioural therapy

1 | INTRODUCTION

Internet-delivered cognitive behavioural therapy (iCBT) is a first-line treatment for mild-to-moderate depression and anxiety (e.g., NICE, 2023b, 2023a). However, effect sizes, particularly within routine care, remain modest, with no increase in magnitude over the past two decades (Moshe et al., 2021). Providing feedback to therapists regarding patients' symptom change over time ('feedback-informed treatment' [FIT]) can improve outcomes by providing clinicians with insight into whether the therapeutic approach benefits the patient (de Jong et al., 2021). Building on this, other work has aimed to provide clinicians with predictions of how likely it is that an individual will respond to treatment, enabling the potential for early adaptations to treatment for likely non-responders. To date, FIT models in routine practice have been based on prediction models that primarily rely on baseline symptom severity to indicate likely treatment response (Webb & Cohen, 2021). Newer models using machine learning methods (such as deep-learning models [DLMs]) have shown superior accuracy (Bone et al., 2021), but have not been tested prospectively in routine clinical care. There is also scarce work examining whether incorporating feedback within iCBT improves treatment or is acceptable to therapists. In the current study, we examine the acceptability and effectiveness of a DLM predicting treatment response to iCBT in a randomised controlled trial conducted in routine care.

1.1 | Feedback-informed treatment

FIT is the process of providing therapists with ongoing insights into client progress during therapy. The kinds of feedback within FIT can differ; for instance, feedback could be client treatment progress in the form of measuring changes in clinical symptoms, or feedback could be based on a client's perception of the therapeutic alliance (Miller et al., 2016). Feedback provides therapists with opportunities to understand whether the client is benefitting, and to adjust the therapeutic approach if needed. The focus in this study is on feedback regarding changes in client clinical symptoms during treatment. Such feedback has been shown to result in better treatment retention and reduced client deterioration, compared with non-feedback-informed therapy, particularly for clients who are not improving (i.e., 'not-on-track' cases; NOT; Lambert et al., 2001; Rognstad et al., 2023). Other work has demonstrated comparable symptom reduction in significantly fewer sessions and lower dropout rates (Janse et al., 2020). Feedback supports therapists' engagement in deliberate practice behaviours, such as case review and goal setting, which are important for developing effective therapists and improving client outcomes (Budesa, 2020; Chow et al., 2015). While

Implications for practice and policy

- Presenting client prediction feedback from a machine learning algorithm is an acceptable and useful tool for therapists providing iCBT.
- Providing client prediction feedback to therapists enhanced client engagement with iCBT in this study and may improve clinical outcomes for not-on-track clients.
- Incorporating client progress feedback tools promises one way to bolster iCBT treatments and therapists providing iCBT could benefit from being trained in the use of such tools.

FIT studies have shown promise of improving treatment, it is also important to note that meta-analyses have identified substantial variation in FIT effectiveness across trials, as well as publication bias towards not publishing null or negative effects (Lambert et al., 2018; Rognstad et al., 2023).

One predominant method of delivering feedback is utilising the expected treatment response model (Finch et al., 2001). This model uses a patient's baseline severity to generate an expected treatment response trajectory (based on normed treatment response data for patients with that baseline severity). Client progress relative to this expected trajectory can then be used as feedback for whether the patient's treatment response is on track. Implementing this model into the delivery of psychological interventions in routine care is feasible and was found to enhance clinical outcomes in one study (Delgadillo et al., 2017, 2018). In a UK routine care setting, 79 therapists were randomised to receive FIT (i.e., expected treatment response predictions) or treatment as usual (TAU; Delgadillo et al., 2018). Results suggested small improvements, particularly for clients predicted to be 'not-on-track'. While promising for delivering feedback-informed therapy, expected treatment response models have several limitations, such as not incorporating patient characteristics beyond baseline severity and a fixed prediction based upon an initial assessment that does not update as new information is collected (Bone et al., 2021). Fortunately, machine learning techniques present one way to address these limitations (Bone et al., 2021; Thieme et al., 2023).

1.2 | Leveraging machine learning to deliver feedback-informed treatment in icbt

Machine learning techniques can offer advantages over other statistical approaches for predicting client change (e.g., logistic

regression and growth curve models), particularly when the number of predictors is large and the associations among the predictors and outcome is nonlinear (Chekroud et al., 2021). Several studies have highlighted potential benefits of using machine learning techniques in mental health treatment, for example, to identify patients at risk of relapse (Lorimer et al., 2021), to predict symptom improvement and remission (Angstman et al., 2017; Pearson et al., 2019; Wallert et al., 2022), and to forecast treatment nonresponse (Kautzky et al., 2018; Perlis, 2013). Importantly, machine learning models that update predictions as more information accumulates over time outperform the traditional expected treatment response model (e.g., area under the curve [AUC] = .80 vs. .70; Bone et al., 2021). An important next step is to investigate whether such models (a) can be successfully implemented in a way that therapists can use and (b) improve treatment outcomes (Boman et al., 2019; Thieme et al., 2023). iCBT settings appear particularly suited to employing such machine learning predictions given their potential to capture large-scale standardised longitudinal data (Chekroud et al., 2021).

To this end, we previously developed (in partnership with Microsoft) a DLM that utilises repeated assessments of client Patient Health Questionnaire-9 (PHQ-9) and Generalized Anxiety Disorder-7 (GAD-7) scores over time in a recurrent neural network to predict with high accuracy (83.90%) whether a patient will show reliable improvement at the end of an 8 week iCBT intervention in routine care (Prasad et al., 2023). In brief, deep learning is a type of neural network technique within machine learning that leverages multiple intermediate 'hidden layers' in between raw data inputs and outputs (Matheny et al., 2023). These hidden layers can apply transformations and model nonlinear relationships between inputs and outputs. A further subclass of these neural network models, known as recurrent neural networks, further incorporates information regarding the non-independence of repeated client symptom measurement. Thus, the model developed in this previous work utilised a neural network with multiple hidden layers and that accounted for the non-independence of time series data to allow for modelling complex nonlinear associations between repeated measurements of symptoms over time and client outcome. Furthermore, predictions from this model are updated as the client completes more PHQ-9 and GAD-7 questionnaires throughout treatment. This model was built using a large dataset ($N = 45,876$) of iCBT clients in a routine care setting and demonstrated generalisability to different iCBT programs, geographies and demographic groups (accuracies >80%; Prasad et al., 2023). Given the success and generalisability of this model in predicting client treatment response, it is critical to evaluate whether it can be implemented within routine care and whether it enhances client engagement and clinical outcomes. Thus, the current study examined the effectiveness and acceptability of implementing our DLM model as a therapist tool in iCBT treatment for depression and anxiety within a stepped-care setting in the NHS in the UK (i.e., Talking Therapies, Step 2).

1.3 | Study hypotheses

The specific hypotheses were:

1. Clients of therapists in the DLM tool group will have greater symptom reductions on the PHQ-9 and GAD-7 (i.e., greater magnitude of symptom change slopes, greater rates of reliable improvement) than clients of therapists in the TAU control group.
2. Clients of therapists in the DLM tool group will have greater engagement with the iCBT platform than clients of therapists in the TAU control group.
3. Therapists will report the DLM tool acceptable to use for supporting clients.

As a secondary analysis, this study also explored whether 'not-on-track' clients whose therapist had access to the DLM tool will have greater engagement and clinical improvements than 'not-on-track' clients whose therapist is in the TAU control group.

2 | METHODS

2.1 | Study design

A cluster randomised controlled trial (ISRCTN18059067) was conducted, in which therapists (hereafter referred to as psychological well-being practitioners [PWPs] to maintain consistency with the occupational terminology used at the service site utilised in this study) providing support to patients utilising iCBT for depression and anxiety were randomised to have access to the DLM tool or to deliver TAU. The NHS England Research Ethics Committee and Health Research Authority granted ethics approval for this study (IRAS Project ID: 299656). A CONSORT checklist for this study is available in the Data S1.

2.2 | Study setting

This study occurred at Berkshire National Health Service (NHS) Foundation Trust Talking Therapies service, which serves a population of 900,000 individuals across six demographically and economically diverse localities (see Palacios et al., 2023, for more information on service site characteristics, procedures and iCBT-I intervention effectiveness). Talking Therapies manage Steps 2 and 3 of the NHS's stepped-care program. Clients in this study came from Step 2, in which iCBT was delivered by a trained PWP as a low-intensity intervention for mild-to-moderate symptomatology. PWPs are typically psychology graduates with further training in delivering low-intensity cognitive behavioural therapy-based interventions (Clark, 2011). Clients with severe symptoms requiring high-intensity therapies are transferred to Step 3, where they are seen by more

experienced licenced therapists for individual face-to-face treatment and are thus not included in this study.

Individuals who contact the Berkshire NHS Foundation Trust Talking Therapies service complete an intake assessment by phone or in person. This intake assessment includes the PHQ-9, GAD-7, Phobia Scale, and Work and Social Adjustment Scale. The assessment determines the level of symptomatology and the optimal allocation within the stepped-care model. When discussing treatment options with service users, the PWPs provide information about the interventions (i.e., nature, content and duration). The PWP and client then arrive at a collaborative decision regarding treatment. Only clients routed to the SilverCould iCBT program were included in the study sample.

2.3 | Recruitment and randomisation

All 74 PWPs who were providing clinical support to clients in NHS Berkshire Talking Therapies at the time of the study launch were invited to participate in the study. These PWPs attended an information session and/or received an information sheet regarding the study. Fifty-five PWPs consented to participate in the study and were randomised via Qualtrics. Randomisation was stratified based on PWP experience, either novice (have supported 1–40 clients using SilverCloud) or experienced (have supported over 40 clients on SilverCloud; threshold defined based upon discussion with the staff at NHS Berkshire). PWPs could withdraw from the study by contacting study personnel who would remove them from the study. Explicit consent from each client was not sought for this study as the research was determined to fall within the rubric of service development and evaluation, and only anonymous patient data were provided from the service site to study researchers for analysis. All trial participation and data collection occurred in 2022, and data collection ended once the planned sample size was approximately reached. No harms or unintended consequences were reported during the study.

2.4 | Treatment

All clients in this study received one of the SilverCloud® by Amwell's® *Space from Depression*, *Space from Anxiety*, or *Space from Depression and Anxiety* iCBT programs. The suitability of SilverCloud and assignment of program are assessed during an initial phone call between a client and their PWP. These iCBT programs (described in detail elsewhere) deliver evidence-based CBT for depression and anxiety, and have demonstrated effectiveness in multiple clinical trials (National Institute for Health and Care Excellence, 2023a, 2023b; Richards et al., 2015, 2020). In the study's routine care site, PWPs provide support to their clients through regular (typically every 10–14 days) 'reviews', with clients receiving six reviews on average (although this may vary depending on client needs). In reviews, the PWP assesses the client's progress, provides feedback based on the client's activities in the program and provides encouragement to promote meaningful engagement with the program. These reviews are typically provided via short (e.g., ~150 words) written messages within the SilverCloud platform, though clients who needed extra support also received phone call reviews. Potential risk to participants was managed in line with routine service procedures.

2.5 | Deep-learning model tool

The DLM tool was co-designed with PWPs and implemented as a widget into the PWP-facing side of the SilverCloud platform (as described in Thieme et al., 2023). PWPs in the DLM tool condition could view a client's predicted probability of achieving reliable improvement (i.e., decrease in PHQ-9 score of 6 or greater or decrease in GAD-7 score of 4 or greater) by the end of treatment. This prediction was presented as text to PWPs in one of five categories: very unlikely to improve, unlikely to improve, unsure, likely to improve and very likely to improve (see Figure S1 for an example). The tool presents separate predictions for reliable improvement on the PHQ-9 and the GAD-7. A prediction is generated if/once a client meets the following requirements: (1) baseline score of 10 or greater on the PHQ-9, or 8 or greater on the GAD-7, and (2) the patient has completed 3 or more PHQ-9 or GAD-7 measures. The prediction then updates dynamically at each review as the client completes more PHQ-9 and GAD-7 assessments. All PWPs randomised to the DLM tool condition attended a training session that detailed how the tool was developed, its accuracy/generalisability, how to interpret tool predictions and a review of different case scenarios with differing predictions (e.g., no prediction is shown, prediction of reliable improvement, prediction of no reliable improvement and unsure). PWPs also completed an interactive exercise using the tool to increase familiarity with the new functionality. Finally, potential concerns of using machine learning tools were discussed, and PWPs were instructed not to share the predictions with their clients (to not bias or influence their treatment).

2.6 | Assessments

2.6.1 | Assessments for clients

Clients completed routine outcome measures (i.e., PHQ-9 and GAD-7) on the SilverCloud platform at each review timepoint.

Patient health Questionnaire-9

The PHQ-9 is a 9-item self-report measure of depression symptoms, with good test–retest reliability ($r = 0.94$) and internal consistency (ICC = 0.88) (Kroenke et al., 2010; Zuithoff et al., 2010). Summary scores range from 0 to 27; higher scores indicate greater symptom severity. In line with NHS Talking Therapies guidelines, a decrease in PHQ-9 scores of 6 or more was used for determining reliable improvement (National Collaborating Centre for Mental Health, 2018).

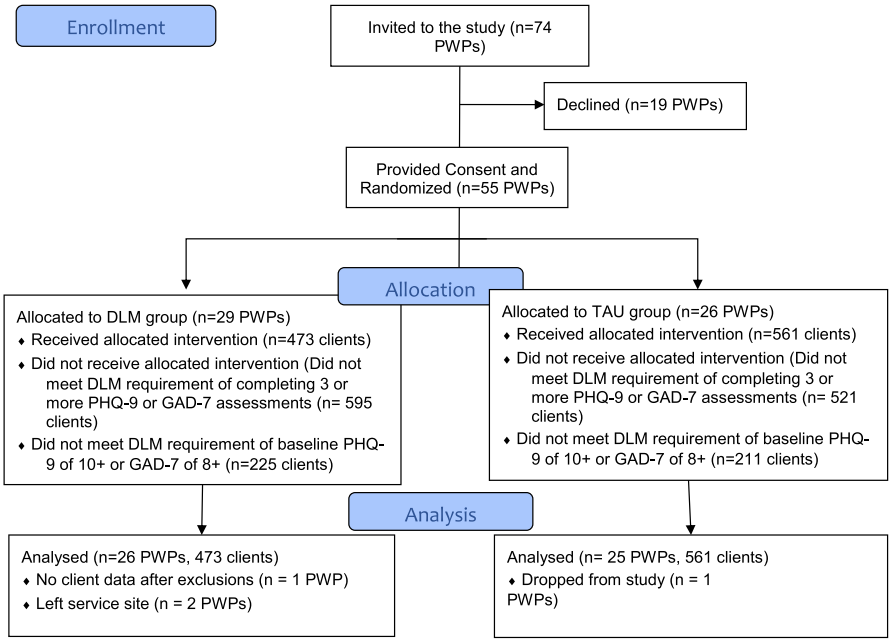

FIGURE 1 Trial psychological well-being practitioners (PWP) and client flow.

Generalized Anxiety Disorder-7

The GAD-7 is a 7-item self-report measure of GAD symptoms, with good internal consistency ($\alpha = .92$) and good convergent validity with other anxiety scales (Spitzer et al., 2006). Summary scores range from 0 to 21, where higher scores indicate greater symptom severity. In line with NHS Talking Therapies guidelines, a decrease in GAD-7 scores of 4 or more was used for determining reliable improvement (National Collaborating Centre for Mental Health, 2018).

Platform engagement

Client platform engagement data are routinely collected through the SilverCloud platform. Aggregate platform engagement metrics were collected during this study: total time on the platform, number of sessions, number of unique tools used, number of pages used and number of activities completed.

Client status of 'not-on-track'

Regardless of a therapist's access to the DLM tool, the DLM algorithm was used to generate predictions for whether a client was predicted to have reliable improvement (i.e., 'on-track') or not ('not-on-track') on the PHQ-9 or GAD-7 for clients in both the DLM and TAU condition. The first prediction was extracted for all clients. If a client had a first prediction for either the PHQ-9 or the GAD-7 indicating that they were very unlikely or unlikely to have reliable improvement, then the client was categorised as not-on-track (0 = on-track, 1 = not-on-track).

2.6.2 | Assessments for PWPs

All consenting PWPs completed a demographic questionnaire at baseline. PWPs in the DLM group additionally completed a set of questions regarding the acceptability of the DLM tool 12 weeks after study initiation. PWPs were provided with a £10 gift card for completion of each assigned time point and a £10 bonus for completing all assigned time points.

PWP DLM tool questions

Assessment of PWP acceptance of the DLM tool used a combination of items adapted from or informed by questionnaires and research on integrating technology into healthcare. Four items on technology acceptance (Nadal et al., 2020) assessed the DLM tool's acceptability. Seven items were generated to assess PWP attitudes towards the DLM tool. These items queried attitudes such as if DLM predictions provided reassurance, if the tool predictions were demotivating and if the DLM predictions were useful for making clinical decisions more quickly. Six items were informed by the Retrospective Analysis of Psychotherapists' Involvement in Deliberate Practice questionnaire (Chow et al., 2015) to assess clinician engagement in activities that improve therapeutic performance. Three of these items assessed PWP behaviour if the client was predicted not to improve while the other three items assessed behaviour if the client was predicted to improve. One last item assessed if the PWP engaged specific behaviours because of the DLM tool (e.g., reflecting on past sessions). Full wording for all questions is provided in the Data S1.

2.7 | Planned statistical power

Statistical power was calculated using PowerUpR, an online app for designing and analysing multilevel randomised controlled trials (Ataneka et al., 2023). This software was used to conduct an a-priori power analysis using effect size and multilevel model parameters identified in Delgadillo et al. (2018) for study planning purposes. Assuming an effect size of $d = 0.23$ with a power of 80%, alpha of 0.05, a PWP level intra-cluster correlation of 0.05, assumed

recruitment of 55 PWPs and a 1:1 randomisation, the software indicated a required client-level sample size of 18. Twenty-five per cent uplift to this target sample was added to ameliorate against attrition. As a result, we aimed to recruit a minimum client sample size of 21, resulting in a target total sample size of 1155 clients (55 PWPs x 21 clients).

2.8 | Sample size and end study power

During the study, 55 PWPs oversaw 2391 clients who completed 7247 assessments during the 8 weeks of treatment. Figure 1 below presents the trial recruitment and exclusion flow for PWPs and clients. After necessary technical exclusions of clients who did not meet the DLM tool prediction requirements (and therefore are not the target clients the tool was developed on and designed for), the final sample consisted of 51 PWPs who oversaw the care of 1034 clients; these clients completed 4549 clinical assessments while using SilverCloud. Thus, on average, a PWP oversaw the care of 20.27 eligible clients during the trial period (SD = 13.58). On average, clients completed 4.40 assessments (SD = 1.39). Of these 51 PWPS, 26 (51%) were randomly assigned to the DLM prediction tool condition and 25 (49%) to the TAU condition. This end sample size and the collected data were leveraged to conduct an end-of-study minimal detectable effect size to check whether end study power was in line with the a prior power analysis. Assuming a power of 80%, alpha of 0.05, 51 PWPs randomised in a 1:1 fashion, 22 clients per PWP, 4 assessments per client, a PWP level intra-cluster correlation of 0.003, a client-level intra-cluster correlation of .62, the end study minimal detectable effect size at the PWP grouping level was $d = .16$. Note that this minimal detectable effect size is nearly the same as the effect size identified in the most up-to-date meta-analysis on FIT RCT trials (Rognstad et al., 2023).

2.9 | Statistical analyses

Data were prepared and analysed using SPSS v28 and Mplus v8; statistical figures were generated in R v4.2 using the tidyr and ggplot2 packages (IBM Corp., 2021; Muthén & Muthén, 2017; R core Team, 2017; Wickham, 2016; Wickham et al., 2017).

To examine the first hypothesis ('the DLM tool will increase symptom reductions'), three-level growth curve models (i.e., assessments nested under clients who were nested under PWPs) were estimated for (a) raw score change and (b) reliable improvement rates in PHQ-9 and GAD-7 outcomes from week 0 to week 8. Growth curve models allowed for random intercepts and slopes at each level to account for individual differences in client initial symptom severity and response to treatment. These models estimate each client's unique trajectory of change in clinical symptoms during the 8-week treatment period and allow for the prediction of these changes over time based on whether the client's supporting PWP was randomised to the DLM or TAU condition. Raw score growth models were linear multilevel models. Reliable improvement was modelled in a logistic multilevel model because it is a binary outcome. Furthermore, because reliable improvement is a binary outcome in which all client values are 0 at baseline and incorporating this uninformative timepoint requires a large departure in modelling a linear time trend (see Figure S3), reliable improvement is modelled beginning at Week 1 for model parsimony and ease of estimation. Missing data within all MPlus analyses were handled using full information maximum likelihood estimation (Hayes & Enders, 2023).

To examine the second hypothesis ('the DLM tool will lead to greater engagement'), two-level multilevel models (clients nested under PWPs) with random intercepts were used to estimate whether PWP assignment to the DLM or TAU condition predicted client usage outcomes (total time on platform, number of sessions, number of unique tools used, number of pages used and number of activities completed).

To examine the exploratory secondary analysis ('does the DLM tool especially benefit clients identified as NOT?'), the multilevel models utilised in Hypotheses 1 and 2 were reconducted but with the addition of a NOT x Condition interaction term in the prediction of each outcome.

Finally, to examine the third hypothesis ('PWPs will report the tool as acceptable'), we descriptively report findings from questionnaires completed by PWPs at 12 weeks regarding tool acceptability.

3 | RESULTS

PWP characteristics are presented in Table S1. Overall, PWPs in this study were predominately female (~90%), ~30 years of age, had 2–3 years of experience as a PWP, and 90% met the study definition of being an experienced PWP (i.e., had previously supported 40 or more clients as a PWP using SilverCloud). As expected with the successful randomisation of PWPs to intervention conditions, these characteristics were approximately equal between the TAU and DLM tool conditions.

3.1 | Did client clinical outcomes improve with PWP access to the DLM tool?

Table 1 presents the raw descriptive statistics for PHQ-9 and GAD-7 scores throughout 8 weeks of treatment (see Figures S2 and S3 for visual depictions). Trends over time displayed a relatively linear trajectory of PHQ-9 and GAD-7 outcomes during treatment. Thus, a linear time trend (i.e., a numerical week since the start of treatment) was used to model change in PHQ-9 and GAD-7 scores over time in multilevel models. Intraclass correlation coefficients were calculated for all outcomes at each level of analysis and are reported in Table S2. These coefficients represent the per cent of variability in the outcome at each level of analysis and demonstrate that most of the variability in PHQ-9 and GAD-7 scores occurs at the

TABLE 1 Raw descriptive statistics for client clinical scores each week.

	TAU		DLM	
	PHQ-9	GAD-7	PHQ-9	GAD-7
Week	M (SD)/*n*	M (SD)/*n*	M (SD)/*n*	M (SD)/*n*
Week 0	13.51 (5.28)/579	13.09 (4.41)/579	13.32 (5.36)/497	13.02 (4.38)/498
Week 1	12.25 (4.95)/133	11.98 (4.63)/134	11.70 (5.54)/137	11.72 (5.77)/137
Week 2	11.50 (5.40)/206	10.81 (5.14)/206	11.06 (5.18)/191	10.41 (4.92) 191
Week 3	11.39 (6.07)/259	11.09 (5.43)/258	10.31 (5.86)/236	10.21 (5.06)/233
Week 4	10.97 (5.36)/269	10.06 (4.99)/269	10.91 (5.97)/255	10.68 (5.43)/255
Week 5	10.46 (6.16)/268	10.05 (5.44)/267	10.37 (5.84)/226	9.92 (5.29)/226
Week 6	10.02 (5.77)/240	9.42 (5.01)/238	9.70 (6.23)/215	9.12 (5.77)/214
Week 7	10.16 (6.00)/231	9.70 (5.42)/229	9.39 (5.75)/204	8.91 (5.24)/204
Week 8	9.04 (5.51)/177	8.86 (5.17)/178	9.66 (6.31)/196	9.37 (5.62)/193

Abbreviations: DLM, deep-learning model tool condition; TAU, treatment as usual condition.

TABLE 2 Multilevel model parameters, variance in parameters and prediction of the parameters by psychological well-being practitioner (PWP) condition.

	Intercept (average value at model time 0)	Slope (average change per week in treatment)
PHQ-9		
Intercept/Slope	Intercept = 13.11*	Slope = −0.52*
Therapist DLM tool condition	*B* = −0.14	*B* = −0.03
Client-level variance	Var = 22.95*	Var = 0.38*
Therapist-level variance	Var = 0.02	Var = 0.00
PHQ-9 Reliable improvement		
Intercept/Slope	Intercept = −2.18*	Slope = 0.15*
Therapist DLM tool condition	*B* = −0.01	*B* = 0.02
Client-level variance	Var = 1.74*	Var = 0.15*
Therapist-level variance	Var = 0.09*	Var = 0.01*
GAD-7		
Intercept/Slope	Intercept = 12.75*	Slope = −0.58*
Therapist DLM tool condition	*B* = −0.15	*B* = 0.02
Client-level variance	Var = 15.14*	Var = 0.31*
Therapist-level variance	Var = 0.05	Var = 0.00
GAD-7 Reliable improvement		
Intercept/Slope	Intercept = −1.29*	Slope = 0.20*
Therapist DLM tool condition	*B* = 0.11	*B* = 0.01
Client-level variance	Var = 1.45*	Var = 0.17*
Therapist-level variance	Var = 0.05*	Var = 0.01*

Abbreviation: *B*, unstandardised effect.

*$p < .05$. Therapist DLM tool condition coded 0 = TAU, 1 = DLM tool. Intercepts parameter for reliable improvement are in logit units.

within-person level (i.e., variation within client over time) and the between-client level (i.e., variation across clients). Less than one per cent of variability in PHQ-9, GAD-7, and platform usage outcomes occurred at the PWP level (i.e., variation across PWPs).

Table 2 presents the key parameters from the three-level multilevel models for each clinical outcome. This table presents the intercept and slope from each model, the variance in each of these parameters, and whether DLM tool condition predicted each. In

terms of PHQ-9 and GAD-7 outcomes, DLM tool condition did not predict baseline score (i.e., the intercept), indicating there were not statistically significant differences in client symptom severity between the DLM tool group and TAU when starting treatment (all *ps* > .05). Furthermore, DLM tool condition did not statistically significantly predict client slopes for clinical outcomes (i.e., magnitude of changes in clinical symptoms from week 0 to week 8; all *ps* > .05). In other words, clients supported by a PWP who had the DLM tool had the same magnitude of changes in PHQ-9 and GAD-7 outcomes over time as clients supported by a PWP who did not have the DLM tool.

3.2 | Did client platform usage increase with PWP access to the DLM tool?

Table 3 presents the raw means and standard deviations of client usage outcomes, and Table 4 presents the key parameters from the two-level multilevel models for each platform usage outcome. There were no statistically significant differences in hours spent and total number of sessions on the platform between the DLM and TAU conditions (both *ps* > .05). However, clients whose PWP was randomised to have access to the DLM tool, on average, used 0.49 more unique platform tools ($d = .14$, $p = .02$), visited 11.31 more pages ($d = .14$, $p = .04$) and completed 13.55 more activities ($d = .14$, $p = .03$) than clients whose PWP was randomised to TAU.

3.3 | Exploration of whether NOT clients benefit more from the DLM tool

The first DLM tool prediction was NOT for 355 of the 561 (63.3%) clients who had PWPs in the TAU condition and for 293 of the 473 (61.9%) clients in the PWP DLM tool condition. Tables 5 and 6 present the results from the multilevel analyses predicting client clinical outcomes and platform engagement after incorporating client NOT prediction and its interaction with PWP DLM tool condition. In terms of clinical outcomes and as expected, clients initially predicted to be NOT had smaller declines in PHQ-9 and GAD-7 scores (B = 0.74, 0.76, respectively) as well as lower probability of reliable improvement on these measures (slope B = −0.43, −0.31, respectively, all *ps* < .05). However, PWP DLM tool condition only statistically significantly interacted with client NOT status for the slope parameter of reliable improvement on the PHQ-9 (B = 0.17, $p = .03$). A visual depiction of this interaction is displayed in Figure 2. Probing the interaction revealed the discrepancy in reliable improvement on the PHQ-9 between NOT and on-track clients was reduced when PWPs had access to the DLM tool vs. TAU (Client NOT B = −.26 vs. −.43, respectively). This interaction revealed that NOT clients with PWPs with the DLM tool were an estimated 5% more likely to achieve reliable improvement than NOT clients with PWPs who did not have the DLM tool. A similar interaction did not emerge for reliable improvement on the GAD-7 (B = 0.02, $p = .75$).

In terms of client engagement, no differences in engagement were observed for clients predicted as NOT vs. those predicted as on-track (all *ps* > .05). PWP DLM tool condition only statistically significantly interacted with client NOT prediction for the number of pages visited (B = 22.09, $p = .02$). A visual depiction of this interaction is provided in Figure S4. Probing this interaction revealed that clients who were on track viewed more pages when their PWP had access to the DLM tool vs. TAU (B = 25.35, $p = .01$), but this difference was not present for clients who were NOT (B = 3.26, $p = .58$).

3.4 | Did PWPs find the DLM tool acceptable for practice?

PWP responses to the DLM tool acceptability questions at 12 weeks of use indicated that, on average, PWPs agreed that the tool was useful (M = 3.70, SD = 0.77), clear and understandable (M = 4.30, SD = 0.72), were comfortable using the tool (M = 4.00, SD = 0.83), and would like to have the tool in the future (M = 3.85. SD = 0.90; see Figure S5).

Attitudes towards the DLM tool tended to be positive in that PWPs felt the tool predictions confirmed their own assessments (M = 3.81, SD = 0.56), and they found the tool reassuring (M = 4.04, SD = 0.85). PWPs, on average, were neutral towards predictions of no client improvement being demotivating (M = 3.07, SD = 0.83) and disagreed that such predictions meant there was nothing they could do to change the outcome (M = 2.19, SD = 0.79). PWPs were neutral towards high caseloads interfering with tool engagement (M = 2.81, SD = 1.18) and that the tool was a useful shortcut for making quicker clinical decisions (M = 3.19, SD = 1.00). PWPs tended to disagree that the tool meant they did not have to gather as much information about the client to determine the next steps (M = 2.48, SD = 1.01). Figures S6 and S7 visually depict these PWP attitudes towards the DLM tool.

When the tool predicted a client not to improve, on average, PWPs tended to agree that they would take a closer look at the client's profile (M = 3.89, SD = 0.64), would spend more time understanding possible blockers (M = 3.78, SD = 0.75) and would feel motivated to think outside the box (M = 3.89, SD = 0.69; see Figure S8). When the tool predicted a client to improve, on average, PWPs tended to disagree that they would focus more time on other clients with greater support needs (M = 2.85, SD = 1.13) or would scale down the level of support (M = 2.67, SD = 1.24), but did agree that they would continue providing the same level of support (M = 4.15, SD = 0.66; see Figure S9).

Responses to deliberate practice effects questions indicated that when having the tool, approximately half of the PWPs spent more time reviewing difficult cases (52.8%), reflecting on past sessions (40.7%), reflecting on what to do in future sessions (59.2%) and engaging in case discussions with a mentor or supervisor (40.7%; see Figure S10). Few PWPs indicated that the tool made them spend more time in case discussions with peers (7.4%) or reading case examples (14.8%).

TABLE 3 Raw descriptive statistics for client usage outcomes.

	TAU ($n = 561$)	DLM ($n = 473$)
	M (SD)	M (SD)
Time on platform (hours)	4.69 (3.57)	5.14 (3.85)
Number of sessions	22.38 (16.32)	23.83 (16.80)
Number of unique tools used	7.43 (3.39)	7.92 (3.58)
Number of pages visited	127.68 (75.41)	138.56 (82.75)
Number of activities completed	136.65 (96.85)	149.99 (101.44)

Abbreviations: DLM, deep-learning model tool condition; TAU, treatment as usual condition.

TABLE 4 Multilevel model parameters, variance in parameters and prediction of the parameters by PWP condition.

	Intercept (i.e., client total during treatment)
Hours on platform	
Intercept	Intercept = 4.74*
Therapist DLM tool condition	$B = 0.40$
Client-level intercept variance	Var = 13.64*
Therapist-level intercept variance	Var = 0.05
Sessions	
Intercept	Intercept = 22.37*
Therapist DLM tool condition	$B = 1.45$
Client-level intercept variance	Var = 272.90*
Therapist-level intercept variance	Var = 0.35
Unique tools used	
Intercept	Intercept = 7.44*
Therapist DLM tool condition	$B = 0.49$*
Client-level intercept variance	Var = 11.68*
Therapist-level intercept variance	Var = 0.03
Pages visited	
Intercept	Intercept = 127.55*
Therapist DLM tool condition	$B = 11.31$*
Client-level intercept variance	Variance = 6184.45*
Therapist-level intercept variance	Var = 24.56
Activities completed	
Intercept	Intercept = 136.57*
Therapist DLM tool condition	$B = 13.55$*
Client-level intercept variance	Var = 9753.77*
Therapist-level intercept variance	Var = 29.33

Abbreviation: *B* unstandardised effect.

*$p < .05$. Therapist DLM Tool condition coded 0 = TAU, 1 = DLM tool.

4 | DISCUSSION

This study examined the effectiveness and acceptability of providing PWPs administering iCBT with a DLM tool that provided feedback on predicted client treatment response. In terms of client outcomes, we did not find overall improved clinical outcomes for clients whose PWP was randomised to access the DLM tool compared with clients whose PWP did not have access to the tool (i.e., TAU). Focusing only on clients predicted to be NOT, however, identified that NOT clients whose PWP had access to the tool were more likely to achieve reliable improvement on the PHQ-9 (vs. TAU). In addition, clients with PWPs with access to the DLM tool had greater platform engagement in terms of increased tool use, pages viewed and completed activities. Finally, based on the high levels of agreement reported across the questions assessing acceptability, we have reasonable assurance that PWPs find the DLM tool acceptable for practice. Encouragingly, we find evidence for increased deliberate practice efforts, a crucial component of FIT models. Altogether, this pattern of findings suggests the DLM tool enhanced iCBT treatment with some clinical benefits for NOT clients.

Considering these findings in the context of the broader literature on FIT, there have been reports of statistically significant improvements in client clinical outcomes when therapists utilise FIT (e.g., Delgadillo et al., 2018; Lambert et al., 2001). However, a recent meta-analysis showed that ~40% of effects from randomised controlled trials examining FIT were null or negative (prior to further accounting for biased non-publication of negative/null effects, see figure 4 in Rognstad et al., 2023). Thus, it is critical to consider when and for whom feedback is beneficial rather than assuming a positive effect in all cases (de Jong et al., 2021; Lambert et al., 2018; Rognstad et al., 2023). One dominant factor producing heterogeneity in findings is whether the study focused on NOT clients, rather than all clients. Identifying and intervening with NOT cases is a key benefit of FIT, mainly as NOT clients are most likely to benefit from changes to therapist behaviour, which is in line with our findings. Future work should continue to focus on these clients when implementing FIT, including conducting trials specifically powered to focus on this subgroup and aiming to understand what feedback and therapist behaviours are most beneficial.

It is also critical to consider the type of treatment in which the feedback was employed. The current study is one of few examining the effects of providing feedback in iCBT in routine care (see Delgadillo et al., 2018, in which iCBT was one of multiple treatment types). While iCBT mirrors face-to-face CBT therapy in many ways, there are critical differences, such as reduced client-therapist interaction time and interaction occurring via a different medium. In this study, PWP communication to the client occurred through bi-weekly written reviews (though sometimes these occurred via telephone). In contrast, prior work finding beneficial feedback effects, such as Delgadillo et al. (2018), utilised predominately synchronous forms of communication and interaction between the therapist and client, such as iCBT with telephone support. Telephone support may provide a better opportunity to utilise feedback to improve client treatment. Therapists may engage in more therapeutic behaviours in phone reviews vs. written reviews, and engagement in such behaviours is correlated with improved client outcomes (Holländare et al., 2016; O'Brien, 2018).

TABLE 5 Multilevel model parameters, variance in parameters and prediction of the parameters by PWP condition, client NOT and condition* NOT interaction.

	Intercept (average value at model time 0)	Slope (average change per week in treatment)
PHQ-9		
Intercept/Slope	Intercept = 12.07*	Slope = −1.02*
Therapist DLM tool condition	*B* = −0.09	*B* = 0.04
Client NOT	*B* = 1.60*	*B* = 0.76*
Therapist DLM tool* Client NOT	N/A	*B* = −0.09
Client-level variance	Var = 22.34*	Var = 0.27*
Therapist-level variance	Var = 0.01	Var = 0.00
PHQ-9 Reliable improvement		
Intercept/Slope	Intercept = −1.36*	Slope = 0.39*
Therapist DLM tool condition	*B* = 0.08	*B* = −0.09
Client NOT	*B* = −1.11*	*B* = −0.43*
Therapist DLM tool* Client NOT	N/A	*B* = 0.17*
Client-level variance	Var = 1.24*	Var = 0.14*
Therapist-level variance	Var = 0.05*	Var = 0.01*
GAD-7		
Intercept/Slope	Intercept = 12.35*	Slope = −1.05*
Therapist DLM tool condition	*B* = −0.13	*B* = −0.09
Client NOT	*B* = 0.62*	*B* = 0.74*
Therapist DLM tool* Client NOT	N/A	*B* = −0.09
Client-level variance	Var = 14.92*	Var = 0.20*
Therapist-level variance	Var = 0.03	Var = 0.00
GAD-7 Reliable improvement		
Intercept/Slope	Intercept = −0.32	Slope = 0.37*
Therapist DLM tool condition	*B* = 0.06	*B* = 0.01
Client NOT	*B* = −1.17*	*B* = −0.31*
Therapist DLM tool* Client NOT	N/A	*B* = 0.02
Client-level variance	Var = 0.57*	Var = 0.11*
Therapist-level variance	Var = 0.02*	Var = 0.01*

Abbreviation: *B*, unstandardised effect.

*$p < .05$. Therapist DLM tool condition coded 0 = TAU, 1 = DLM tool, client NOT coded 0 = on-track, 1 = not-on-track. Intercept parameters for reliable improvement are in logit units.

The mode of bi-weekly written reviews in the current study may not provide sufficient opportunity for a PWP to utilise feedback to significantly alter client outcomes in comparison with telephone support or face-to-face therapy. It is also important to note that therapist-level variance in engagement and clinical outcomes was minimal in this study (<1%), suggesting a large degree of similarity/fidelity across therapists and perhaps limited room for the DLM tool to improve therapeutic processes.

In contrast to distal clinical outcomes, feedback tools may have a greater impact on more proximal treatment factors, such as client engagement. We found that clients whose PWP had access to the DLM tool used more unique platform tools, visited more pages and

TABLE 6 Multilevel model parameters, variance in parameters and prediction of the parameters by PWP condition, client NOT and condition* NOT interaction.

	Intercept (i.e., client total during treatment)
Hours on platform	
Intercept	Intercept = 4.65*
Therapist DLM tool condition	*B* = 0.91*
Client NOT	*B* = 0.23
Therapist DLM tool* Client NOT	*B* = −0.76
Client-level variance	Var = 13.68*
Therapist-level variance	Var = 0.02
Sessions	
Intercept	Intercept = 22.02*
Therapist DLM tool condition	*B* = 2.58
Client NOT	*B* = 1.03
Therapist DLM tool* Client NOT	*B* = −1.70
Client-level variance	Var = 274.95*
Therapist-level variance	Var = 0.22
Unique tools used	
Intercept	Intercept = 7.32*
Therapist DLM tool condition	*B* = 0.75*
Client NOT	*B* = 0.30
Therapist DLM tool* Client NOT	*B* = −0.40
Client-level variance	Var = 11.62*
Therapist-level variance	Var = 0.03
Pages visited	
Intercept	Intercept = 124.21*
Therapist DLM tool condition	*B* = 25.35*
Client NOT	*B* = 8.59
Therapist DLM tool* Client NOT	*B* = −22.09*
Client-level variance	Var = 6133.31*
Therapist-level variance	Var = 22.24
Activities completed	
Intercept	Intercept = 138.67*
Therapist DLM tool condition	*B* = 14.00
Client NOT	*B* = 0.13
Therapist DLM tool* Client NOT	*B* = −0.24
Client-level variance	Var = 9768.41
Therapist-level variance	Var = 35.57

Abbreviations; *B*, unstandardised effect.

*$p < .05$. Therapist DLM tool condition coded 0 = TAU, 1 = DLM tool, Client NOT coded 0 = on-track, 1 = not-on-track.

completed more activities, but did not spend more time on the platform or have more sessions. This pattern of effects suggests that PWPs with access to the DLM tool may be engaging in deliberate practice by considering tool feedback and directing the client to tools, pages or activities. It seems more plausible that access to the tool would influence client engagement (vs. client clinical outcomes) because PWPs can more easily advise and direct clients to use such platform resources within reviews. In theory, engagement with the active ingredients of treatment should be a precursor to a clinical outcome. Therefore, feedback tools promoting engagement is a welcome result. Previous work has demonstrated that there are likely thresholds, or ranges of engagement, that lead to positive clinical outcomes (Cumpanasoiu et al., 2023; Enrique et al., 2019).

Finally, our findings support that PWP attitudes towards the tool were positive. PWPs generally agreed that the tool was useful, understandable, and that they would like to have the tool in the future. Responses also demonstrated evidence for increased deliberate practice, a central ingredient in FIT models. For instance, PWPs reported that the tool tended to prompt closer examination and planning for clients, particularly clients predicted to not improve. Such behaviours are critical to the development of effective therapists and enhancing client outcomes (Chow et al., 2015). Promoting deliberate practice and flagging these not-on-track cases may be why PWPs viewed the tool positively, similar to findings from Lambert et al. (2001). Despite seemingly putting more effort and planning towards clients predicted to not improve, PWPs were neutral or did not agree that the tool feedback led them to scale down support for or focus away from 'on-track' cases. Together, this suggests that PWPs scaled up effort towards not-on-track clients but did not scale down effort towards on-track clients. This contrasts with prior theorising and findings that feedback should lead to reallocating limited resources away from on-track cases towards not-on-track cases (Lambert et al., 2001). However, it is worth noting that (a) PWPs may be reluctant to report that they are purposefully spending less time and effort on any patients, (b) PWP time/effort resources in our study may not be maximised to the point of having to prioritise some clients at the cost of others, (c) if particular clients are improving then PWPs may not want to reduce efforts in case client progress deteriorates, or (d) PWPs are allocated equal time for each client review, so it might be about 'how' they choose to use that time differently across clients.

5 | STRENGTHS AND LIMITATIONS

This study has key strengths worth highlighting. First, the DLM model used was built, validated and now implemented, within routine care using data from the SilverCloud iCBT depression and anxiety programs and therefore the findings come from an ecologically valid setting. Second, this study provides insights into experiences and effects of the practical implementation of an ML algorithm within psychotherapy to bolster FIT, a key next step in leveraging recent advances in ML. Implementing ML within a digitally delivered treatment is notable because such settings benefit from routine and automated data collection across engagement and outcome measures over time. Continued work in this area will be key to realising the promised advances that ML has to offer.

In addition to these strengths, results of this study should be interpreted while considering several limitations. First, given the nature of the intervention, therapists were not blind to whether

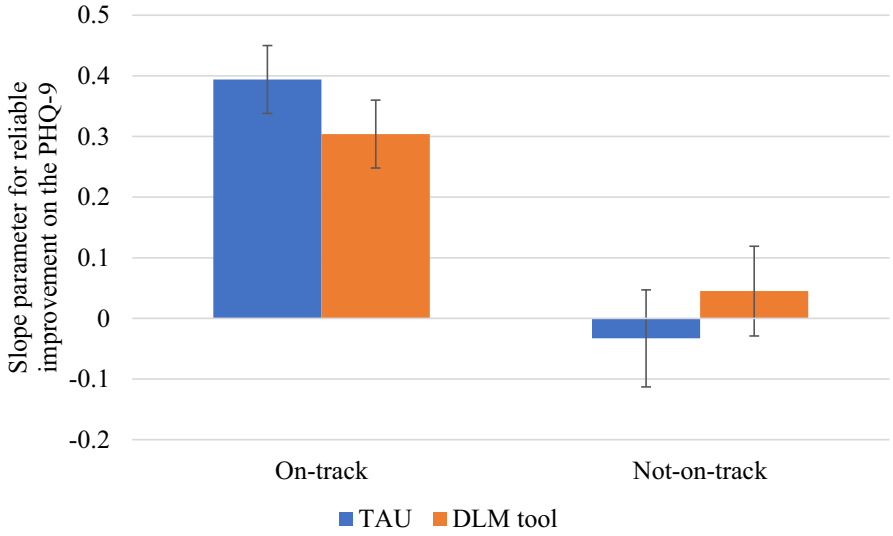

FIGURE 2 Visual depiction of the interaction between client NOT and psychological well-being practitioners (PWP) deep-learning model (DLM) tool condition on the change slope for client reliable improvement on the Patient Health Questionnaire-9. Bars represent standard error. DLM, deep-learning model tool condition; TAU, Treatment as usual condition.

they were in the DLM tool or TAU condition. Knowing they were participating and being monitored in the study, therapists in either group may have been particularly sensitive to client progress and put more effort than usual towards such clients. Second, the DLM tool is not applicable to all clients presenting at a given service site because it requires that a client meets the clinical threshold on the PHQ-9 or GAD-7 at baseline and for that client to complete three assessments in order to generate a prediction. Thus, clients for whom it can generate predictions have already been in treatment. It may be that feedback is most useful in the beginning stages of iCBT treatment. For instance, in our study, there was a notable client drop-off during the first 2–3 weeks of treatment. However, given the DLM tool requirements for prediction generation, a separate algorithm and tool would need to be designed for this use case. Feedback tools that can more quickly identify clients who are not responding to treatment or are at risk of leaving treatment early may prevent such dropout and be even more effective. Third, therapist reports of DLM tool acceptability may be biased by social desirability to respond positively to the DLM tool. Fourth, this is a single-site study; it is therefore unknown if findings would generalise to other services, though prior work has supported that the socio-demographic and clinical characteristics of the service site study is generally representative of the overall IAPT population (Richards et al., 2020). Sixth, given that not all PWPs who were invited to participate completed this study, it is possible this study only recruited a subgroup of PWPs who are highly motivated to use feedback tools, which, if true, could bias ratings of tool acceptability. Fifth, the finding regarding improved PHQ-9 reliable improvement rates for NOT client warrants replication given that parallel findings did not emerge for PHQ-9 change score nor the GAD-7. Last, all findings in this study are limited to the study setup of iCBT treatment wherein PWPs at the NHS Berkshire site typically provided bi-weekly written (though sometimes phone call) reviews to presenting clients. Altogether, based on the promise of the findings and their associated limitations, future work should replicate findings to further enhance confidence in the results.

6 | CONCLUSION

The results of this randomised controlled trial support that the DLM tool was useful and beneficial for therapists and increased client engagement. Clinical benefits were specific to enhanced PHQ-9 reliable improvement rates for NOT clients. Findings support the idea that ML can provide FIT to bolster outcomes for clients predicted NOT to improve while identifying future areas of research and improvement. Particularly, future research studies examining FIT should focus on NOT groups and continue to examine contexts in which ML feedback tools can generate improvements in client outcomes in iCBT.

ACKNOWLEDGEMENTS

We want to thank the Berkshire NHS Foundation Trust as a host site for the research, the PWPS at Berkshire, who participated as participants, and Sarah Sollesse and all other Berkshire site staff who helped coordinate and implement this study. Additionally, we would like to thank our colleagues who collaborated on developing the DLM, including Niranjani Prasad, Isabel Chien, Usman Munir, Ryutaro Tanno, Hannah Richardson, Gavin Doherty (TCD), Aditya Nori, Danielle Belgrave, and Anja Thieme at Microsoft and the project Talia.

FUNDING INFORMATION

This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

CONFLICT OF INTEREST STATEMENT

GH, KSY, DD, AE, DK and DR are employees and shareholders of Amwell, a company that SilverCloud Health has been a subsidiary of since 2021. CC and JP were employees at SilverCloud/Amwell while this work was in progress.

DATA AVAILABILITY STATEMENT

The data for this study are available on request from the corresponding author. The data are not publicly available due to privacy

restrictions, being data that relates to third-party service provision and evaluation, which was shared solely for the intended purpose of the objectives outlined in this paper.

ORCID

Garrett C. Hisler https://orcid.org/0000-0001-7099-8417

REFERENCES

Angstman, K. B., Garrison, G. M., Gonzalez, C. A., Cozine, D. W., Cozine, E. W., & Katzelnick, D. J. (2017). Prediction of primary care depression outcomes at six months: Validation of DOC-6 ©. *Journal of the American Board of Family Medicine*, *30*(3), 281–287. https://doi.org/10.3122/jabfm.2017.03.160313

Ataneka, A., Kelcey, B., Dong, N., Bulus, M., & Bai, F. (2023). PowerUp R Shiny App (v. 0.9) manual. https://powerupr.shinyapps.io/index/

Boman, M., Ben Abdesslem, F., Forsell, E., Gillblad, D., Görnerup, O., Isacsson, N., Sahlgren, M., & Kaldo, V. (2019). Learning machines in internet-delivered psychological treatment. *Progress in Artificial Intelligence*, *8*(4), 475–485. https://doi.org/10.1007/s13748-019-00192-0

Bone, C., Simmonds-Buckley, M., Thwaites, R., Sandford, D., Merzhvynska, M., Rubel, J., Deisenhofer, A.-K., Lutz, W., & Delgadillo, J. (2021). Dynamic prediction of psychological treatment outcomes: Development and validation of a prediction model using routinely collected symptom data. *The Lancet Digital Health*, *3*(4), e231–e240. https://doi.org/10.1016/S2589-7500(21)00018-2

Budesa, Z. (2020). Feedback-informed-treatment: A deliberate approach to responsible practice. *International Journal on Responsibility*, *3*(2), 8.

Chekroud, A. M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., Cohen, Z., Belgrave, D., DeRubeis, R., Iniesta, R., Dwyer, D., & Choi, K. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*, *20*(2), 154–170. https://doi.org/10.1002/wps.20882

Chow, D. L., Miller, S. D., Seidel, J. A., Kane, R. T., Thornton, J. A., & Andrews, W. P. (2015). The role of deliberate practice in the development of highly effective psychotherapists. *Psychotherapy (Chicago, Ill.)*, *52*(3), 337–345. https://doi.org/10.1037/pst0000015

Clark, D. M. (2011). Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: The IAPT experience. *International Review of Psychiatry (Abingdon, England)*, *23*(4), 318–327. https://doi.org/10.3109/09540261.2011.606803

Cumpanasoiu, D. C., Enrique, A., Palacios, J. E., Duffy, D., McNamara, S., & Richards, D. (2023). Trajectories of symptoms in digital interventions for depression and anxiety using routine outcome monitoring data: Secondary analysis study. *JMIR mHealth and uHealth*, *11*, e41815. https://doi.org/10.2196/41815

de Jong, K., Conijn, J. M., Gallagher, R. A. V., Reshetnikova, A. S., Heij, M., & Lutz, M. C. (2021). Using progress feedback to improve outcomes and reduce drop-out, treatment duration, and deterioration: A multilevel meta-analysis. *Clinical Psychology Review*, *85*, 102002. https://doi.org/10.1016/j.cpr.2021.102002

Delgadillo, J., de Jong, K., Lucock, M., Lutz, W., Rubel, J., Gilbody, S., Ali, S., Aguirre, E., Appleton, M., Nevin, J., O'Hayon, H., Patel, U., Sainty, A., Spencer, P., & McMillan, D. (2018). Feedback-informed treatment versus usual psychological treatment for depression and anxiety: A multisite, open-label, cluster randomised controlled trial. *The Lancet Psychiatry*, *5*(7), 564–572. https://doi.org/10.1016/s2215-0366(18)30162-7

Delgadillo, J., Overend, K., Lucock, M., Groom, M., Kirby, N., McMillan, D., Gilbody, S., Lutz, W., Rubel, J. A., & de Jong, K. (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

Enrique, A., Palacios, J. E., Ryan, H., & Richards, D. (2019). 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*, *21*(8), e12775. Https://Www.Jmir.Org/2019/8/E12775

Finch, A. E., Lambert, M. J., & Schaalje, B. G. (2001). Psychotherapy quality control: The statistical generation of expected recovery curves for integration into an early warning system. *Clinical Psychology & Psychotherapy*, *8*(4), 231–242. https://doi.org/10.1002/cpp.286

Hayes, T., & Enders, C. K. (2023). Maximum likelihood and multiple imputation missing data handling: How they work, and how to make them work in practice.

Holländare, F., Gustafsson, S. A., Berglind, M., Grape, F., Carlbring, P., Andersson, G., Hadjistavropoulos, H., & Tillfors, M. (2016). Therapist behaviours in internet-based cognitive behaviour therapy (ICBT) for depressive symptoms. *Internet Interventions*, *3*, 1–7. https://doi.org/10.1016/j.invent.2015.11.002

IBM Corp. (2021). *IBM SPSS statistics for windows, version 28.0*. IBM Corp.

Janse, P. D., de Jong, K., Veerkamp, C., van Dijk, M. K., Hutschemaekers, G. J. M., & Verbraak, M. J. P. M. (2020). The effect of feedback-informed cognitive behavioral therapy on treatment outcome: A randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *88*(9), 818–828. https://doi.org/10.1037/ccp0000549

Kautzky, A., Dold, M., Bartova, L., Spies, M., Vanicek, T., Souery, D., Montgomery, S., Mendlewicz, J., Zohar, J., Fabbri, C., Serretti, A., Lanzenberger, R., & Kasper, S. (2018). Refining prediction in treatment-resistant depression: Results of machine learning analyses in the TRD III sample. *Journal of Clinical Psychiatry*, *79*(1), 16m11385. https://doi.org/10.4088/JCP.16m11385

Kroenke, K., Spitzer, R. L., Williams, J. B. W., & Löwe, B. (2010). The patient health questionnaire somatic, anxiety, and depressive symptom scales: A systematic review. *General Hospital Psychiatry*, *32*(4), 345–359. https://doi.org/10.1016/j.genhosppsych.2010.03.006

Lambert, M. J., Whipple, J. L., & Kleinstäuber, M. (2018). Collecting and delivering progress feedback: A meta-analysis of routine outcome monitoring. *Psychotherapy (Chicago, Ill.)*, *55*(4), 520–537. https://doi.org/10.1037/pst0000167

Lambert, M. J., Whipple, J. L., Smart, D. W., Vermeersch, D. A., Nielsen, S. L., & Hawkins, E. J. (2001). The effects of providing therapists with feedback on patient Progress during psychotherapy: Are outcomes enhanced? *Psychotherapy Research*, *11*(1), 49–68. https://doi.org/10.1080/713663852

Lorimer, B., Delgadillo, J., Kellett, S., & Lawrence, J. (2021). Dynamic prediction and identification of cases at risk of relapse following completion of low-intensity cognitive behavioural therapy. *Psychotherapy Research*, *31*, 19–32.

Matheny, M. E., Ohno-Machado, L., Davis, S. E., & Nemati, S. (2023). Data-driven approaches to generating knowledge: Machine learning, artificial intelligence, and predictive modeling. In *Clinical decision support and beyond* (pp. 217–255). Academic Press.

Miller, S. D., Bargmann, S., Chow, D., Seidel, J., & Maeschalck, C. (2016). Feedback-informed treatment (FIT): Improving the outcome of psychotherapy one person at a time. In *Quality improvement in behavioral health* (pp. 247–262). Springer International Publishing.

Moshe, I., Terhorst, Y., Philippi, P., Domhardt, M., Cuijpers, P., Cristea, I., Pulkki-Råback, L., Baumeister, H., & Sander, L. B. (2021). Digital interventions for the treatment of depression: A meta-analytic review. *Psychological Bulletin*, *147*(8), 749–786. https://doi.org/10.1037/bul0000334

Muthén, L. K., & Muthén, B. O. (1998-2017).). *Mplus user's guide. Eighth edition*. Muthén & Muthén.

Nadal, C., Sas, C., & Doherty, G. (2020). Technology acceptance in mobile health: scoping review of definitions, models, and measurement. *Journal of Medical Internet Research*, *22*(7), e17256.

National Collaborating Centre for Mental Health. (2018). The improving access to psychological therapies manual appendices and helpful resources.

National Institute for Health and Care Excellence. (2023a). Digitally enabled therapies for adults with anxiety disorders: Early value assessment (HTE9.

National Institute for Health and Care Excellence. (2023b). Digitally enabled therapies for adults with depression: Early value assessment (HTE8).

O'Brien, E. (2018). *Therapist behaviours, the working alliance and clinician experience in iCBT for depression and anxiety*. Trinity College Dublin.

Palacios, J., Adegoke, A., Wogan, R., Duffy, D., Earley, C., Eilert, N., Enrique, A., Sollesse, S., Chapman, J., & Richards, D. (2023). Comparison of outcomes across low-intensity psychological interventions for depression and anxiety within a stepped-care setting: A naturalistic cohort study using propensity score modelling. *British Journal of Psychology*, *114*(2), 299–314.

Pearson, R., Pisner, D., Meyer, B., Shumake, J., & Beevers, C. G. (2019). A machine learning ensemble to predict treatment outcomes following an internet intervention for depression. *Psychological Medicine*, *49*(14), 2330–2341. https://doi.org/10.1017/S003329171800315X

Perlis, R. H. (2013). A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. *Biological Psychiatry*, *74*(1), 7–14. https://doi.org/10.1016/j.biopsych.2012.12.007

Prasad, N., Chien, I., Regan, T., Enrique, A., Palacios, J., Keegan, D., Munir, U., Tanno, R., Murfet, H., Nori, A., Richards, D., Doherty, G., Belgrave, D., & Thieme, A. (2023). Deep learning for the prediction of clinical outcomes in internet-delivered CBT for depression and anxiety. *PLoS One*, e0272685.

R core Team. (2017). R: A language and environment for statistical computing.

Richards, D., Enrique, A., Eilert, N., Franklin, M., Palacios, J., Duffy, D., Earley, C., Chapman, J., Jell, G., Sollesse, S., & Timulak, L. (2020). A pragmatic randomized waitlist-controlled effectiveness and cost-effectiveness trial of digital interventions for depression and anxiety. *npj Digital Medicine*, *3*(1), 85. https://doi.org/10.1038/s41746-020-0293-8

Richards, D., Timulak, L., O'brien, E., Hayes, C., Vigano, N., Sharry, J., & Doherty, G. (2015). A randomized controlled trial of an internet-delivered treatment: Its potential as a low-intensity community intervention for adults with symptoms of depression. *Behaviour Research and Therapy*, *75*, 20–31. https://doi.org/10.1186/ISRCTN03704676

Rognstad, K., Wentzel-Larsen, T., Neumer, S.-P., & Kjøbli, J. (2023). A systematic review and meta-analysis of measurement feedback Systems in Treatment for common mental health disorders. *Administration and Policy in Mental Health and Mental Health Services Research*, *50*(2), 269–282. https://doi.org/10.1007/s10488-022-01236-9

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097.

Thieme, A., Hanratty, M., Lyons, M., Palacios, J., Marques, R. F., Morrison, C., & Doherty, G. (2023). Designing human-centered AI for mental health: Developing clinically relevant applications for online CBT treatment. *ACM Transactions on Computer-Human Interaction*, *30*(2), 1–50. https://doi.org/10.1145/3564752

Wallert, J., Boberg, J., Kaldo, V., Mataix-Cols, D., Flygare, O., Crowley, J. J., Halvorsen, M., Ben Abdesslem, F., Boman, M., Andersson, E., Hentati Isacsson, N., Ivanova, E., & Rück, C. (2022). Predicting remission after internet-delivered psychotherapy in patients with depression using machine learning and multi-modal data. *Translational Psychiatry*, *12*(1), 357. https://doi.org/10.1038/s41398-022-02133-3

Webb, C. A., & Cohen, Z. D. (2021). Progress towards clinically informative data-driven decision support tools in psychotherapy. *The Lancet Digital Health*, *3*(4), e207–e208. https://doi.org/10.1016/S2589-7500(21)00042-X

Wickham, H., Vaughan, D., & Girlich, M. (2024). tidyr: Tidy messy data. R package version 1.3.1. https://github.com/tidyverse/tidyr; https://tidyr.tidyverse.org

Wickham, H., Vaughan, D., & Girlich, M. (2017). Tidyr: Tidy messy data. Easily tidy data with'spread'and'gather ()'Functions. https://CRAN.R-project.org/package=tidyr

Zuithoff, N. P. A., Vergouwe, Y., King, M., Nazareth, I., van Wezep, M. J., Moons, K. G. M., & Geerlings, M. I. (2010). The patient health Questionnaire-9 for detection of major depressive disorder in primary care: Consequences of current thresholds in a crosssectional study. *BMC Family Practice*, *11*, 98. https://doi.org/10.1186/1471-2296-11-98

AUTHOR BIOGRAPHIES

Garrett C. Hisler is a Digital Health Scientist at Amwell.

Katherine S. Young is a Principal Digital Health Scientist at Amwell.

Diana Catalina Cumpanasoiu was a Digital Health Scientist at Amwell during this work.

Jorge E. Palacios was a Senior Manager Digital Health Scientist at Amwell during this work.

Daniel Duffy is a Senior Manager Digital Health Scientist at Amwell.

Angel Enrique is a Senior Manager Digital Health Scientist at Amwell.

Dessie Keegan is a Senior Manager Software Engineer at Amwell.

Derek Richards is the Head of Research at Amwell.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Hisler, G. C., Young, K. S., Cumpanasoiu, D. C., Palacios, J. E., Duffy, D., Enrique, A., Keegan, D., & Richards, D. (2025). Incorporating a deep-learning client outcome prediction tool as feedback in supported internet-delivered cognitive behavioural therapy for depression and anxiety: A randomised controlled trial within routine clinical practice. *Counselling and Psychotherapy Research*, *25*, e12771. https://doi.org/10.1002/capr.12771

QUESTIONNAIRE
C8(25)
Incorporating a deep-learning client outcome prediction tool as feedback in supported internet-delivered cognitive behavioural therapy for depression and anxiety: A randomised controlled trial within routine clinical practice

INSTRUCTIONS

- Read through the article and answer the multiple-choice questions provided below.
- **Some questions may have more than one correct answer;** in which case you must mark **all** the correct answers.

Introduction to Feedback-Informed Treatment (FIT) and Machine Learning

Question 1: Traditional expected treatment response models primarily rely on what factor to generate a client's expected treatment trajectory?

- **A:** The client's perception of the therapeutic alliance
- **B:** The number of sessions the client has completed
- **C:** A patient's baseline severity
- **D:** The therapist's years of experience

Question 2: What is identified as a key limitation of traditional expected treatment response models compared to newer machine learning techniques?

- **A:** Difficult to implement in routine care settings
- **B:** They provide a fixed prediction that does not update with new information
- **C:** They are less feasible to implement than machine learning models
- **D:** They only work for mild anxiety and depression

Question 3: In addition to clinical outcomes, what was a specific hypothesis the study aimed to test regarding the impact of the deep learning model (DLM) tool?

- **A:** That therapists using the tool would report lower levels of burnout
- **B:** That clients of therapists in the DLM tool group will have greater engagement with the internet-delivered cognitive behavioural therapy (iCBT) platform than clients of therapists in the treatment as usual (TAU) group
- **C:** That the DLM tool would be more cost-effective than treatment as usual
- **D:** That clients would prefer therapists who used the DLM tool

Study Methods

Question 4: In the study, what criteria did a client have to meet for the DLM tool to generate a prediction?

- **A:** A baseline score of 10 or greater on the PHQ-9 or 8 or greater on the GAD-7, and completion of 3 or more of these measures
- **B:** The client had to complete at least four iCBT modules and be in treatment for 6 weeks
- **C:** The therapist had to manually request a prediction and the client had to give explicit consent
- **D:** A baseline PHQ-9 and GAD-7 score below the clinical threshold and completion of at least one assessment

Study Results

Question 5: What was the main finding from the multilevel growth curve models regarding the DLM tool's impact on overall client clinical outcomes?

- **A:** Clients supported by a psychological well-being practitioner (PWP) with the DLM tool had a greater magnitude of change in GAD-7 scores but not PHQ-9 scores
- **B:** Clients in the TAU group had significantly better clinical outcomes than the DLM tool group
- **C:** The DLM tool did not statistically significantly predict the magnitude of changes in clinical symptoms over time
- **D:** There were statistically significant improvements for all clinical outcomes in the DLM tool group compared to TAU

Question 6: Which one of the following statements accurately describes the effect of the DLM tool on client platform engagement?

- **A:** There were no statistically significant differences in <u>any</u> platform usage metrics between the DLM and TAU groups
- **B:** Clients in the DLM tool group used more unique platform tools, visited more pages, and completed more activities
- **C:** Clients in the DLM tool group spent significantly more hours on the platform and had more sessions
- **D:** The only significant difference was that clients in the DLM group completed more sessions

Question 7: In the secondary analysis, what was the specific clinical benefit for clients predicted to be 'not-on-track' (NOT) whose therapist had access to the DLM tool?

- **A:** They had smaller declines in GAD-7 scores than on-track clients
- **B:** They were significantly more likely to achieve reliable improvement on the GAD-7
- **C:** Their probability of reliable improvement on the PHQ-9 was reduced
- **D:** They were an estimated 5% more likely to achieve reliable improvement on the PHQ-9 than 'not-on-track' (NOT) clients in the TAU group

Question 8: How did Psychological Well-being Practitioners (PWPs) generally rate the acceptability of the DLM tool?

- **A:** PWPs agreed that the tool was useful, clear, understandable, and that they were comfortable using it
- **B:** PWPs were neutral about the tool's usefulness and disagreed that it was clear and understandable
- **C:** PWPs reported that high caseloads interfered with their ability to engage with the tool
- **D:** PWPs found the predictions to be demotivating and disagreed that the tool was reassuring

Question 9: When the DLM tool predicted a client would *not* improve, what was a common behavioral response reported by PWPs?

- **A:** They would scale down the level of support for that client
- **B:** They disagreed that they would spend more time trying to understand blockers
- **C:** They tended to agree that they would take a closer look at the client's profile, would spend more time understanding possible blockers, and felt motivated to think outside the box
- **D:** They felt there was nothing they could do to change the outcome

Question 10: When the DLM tool predicted a client *would* improve, how did PWPs report this influenced their level of support?

- **A:** They would focus more time on other clients with greater support needs
- **B:** They agreed that they would continue providing the same level of support
- **C:** They agreed they would scale down the level of support for that client
- **D:** They would use it as a shortcut and gather less information about the client

Discussion, limitations and conclusion

Question 11: Which of the following statements are **TRUE?**

- **A:** PWPs find the tool acceptable for practice
- **B:** There is evidence for increased deliberate practice efforts, a crucial component of FIT models
- **C:** Altogether, this pattern of findings suggests the DLM tool enhanced iCBT treatment with some clinical benefits for NOT clients
- **D:** None of the above

Question 12: The discussion highlights one dominant factor that contributes to the varied findings across different FIT studies. What is this factor?

- **A:** Whether the study focused on 'not-on-track' (NOT) clients versus all clients
- **B:** The specific type of machine learning model used for predictions
- **C:** The geographical location and culture of the service site
- **D:** The professional background of the therapists (e.g., PWP vs. licensed therapist)

Question 13: What is a key difference in the iCBT context of this study that may have influenced the results compared to other successful FIT trials?

- **A:** The iCBT programs used were not based on evidence-based principles
- **B:** The therapists had significantly less training than in other studies
- **C:** Reduced client-therapist interaction time and communication primarily through written reviews
- **D:** The use of the PHQ-9 and GAD-7, which are not sensitive to change

Question 14: What does the article identify as a key strength of the study?

- **A:** Therapists were successfully blinded to the condition they were assigned to
- **B:** The DLM was built, validated, and implemented within an ecologically valid, routine care setting
- **C:** The study was a multi-site trial, enhancing the generalisability of the findings
- **D:** The DLM tool did not require any prior client data to generate an initial prediction

Question 15: What is the overall conclusion of the randomized controlled trial?

- **A:** The DLM tool increased client engagement and was beneficial for therapists, with clinical benefits specific to 'not-on-track' clients
- **B:** The DLM tool was not acceptable to therapists and its use should be reconsidered in iCBT settings
- **C:** The DLM tool had no effect on client engagement and failed to produce any clinical benefits
- **D:** The DLM tool provided significant clinical benefits for all clients, regardless of their predicted track

THE END

FOCUS ON HEALTH
www.foh-cpd.co.za
Cell: 074 230 3874
Tel: 012 653 0133 / 2373 / 2394
Mon-Fri: 08:00-16:00

400 Theuns van Niekerk Street
Wierda Park
0157

PO Box 71
Wierda Park
0149

MEMBER INFO

(Complete sections marked with an asterisk *)

*FOH number	*HPCSA number

ID number

Personal detail changes

I hereby declare that the completion of this document is my own effort without any assistance.

SIGNED: DATE:

This activity is accredited for **THREE (3) CLINICAL** CEU'S. You have to achieve 70% to pass this activity.

ANSWER SHEET

C8(25)

Incorporating a deep-learning client outcome prediction tool as feedback in supported internet-delivered cognitive behavioural therapy for depression and anxiety: A randomised controlled trial within routine clinical practice

***Time spent on activity hour min**

	A	B	C	D	E		A	B	C	D	E
1						8					
2						9					
3						10					
4						11					
5						12					
6						13					
7						14					
						15					

How relevant was the content of this activity to your scope of practice?			
Not at all	Somewhat	Mostly	Extremely

RETURN ANSWER SHEET TO

Online members, complete your answer sheet online

CPD PORTAL: http://foh-cpd.co.za/cpd-online/cpd-user (Username: FOH number & Password: ID)

(Results are available immediately)

Email or **hard copy members**, complete your answer sheet online or return it to:

WHATSAPP: 074 230 3874 **EMAIL:** info@foh-cpd.co.za

(Results will be sent via email)

You will receive a confirmation of receipt SMS within 12-24hours, if not received please send again.

Have a suggestion or comment?

Send it to us on **EMAIL or WHATSAPP**