Game-based digital intervention for neurocognitive training in major depressive disorder: a randomized double-blinded comparator-controlled clinical trial

J. Matias Palva[1,2,3,4,*,§], Joonas J. Juvonen[1,4,*], Lauri Lukka[1,4], Maria Vesterinen[1,2], Antti Salonen[1,4], Vilma-Reetta Bergman[1], Paula Partanen[1,2], Lauri Pohjola[1,4], Juhani Kolehmainen[1,4], Monika Meimer[5], Xiaosi Gu[6,7,8], Hanna Renvall[1,9], Pekka Jylhä[10], Erkki Isometsä E[10], Satu Palva[2,3]

1. *Department of Neuroscience and Biomedical Engineering, Aalto University, Finland*
2. *Neuroscience Center, Helsinki Institute of Life Science, University of Helsinki, Finland*
3. *School of Psychology and Neuroscience, University of Glasgow, United Kingdom*
4. *Soihtu DTx Ltd, Espoo, Finland*
5. *Department of Psychiatry, Turku University Hospital, Finland*
6. *Department of Psychiatry, Yale School of Medicine, New Haven, CT, USA*
7. *Department of Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, CT, USA*
8. *Center for Brain and Mind Health, Yale School of Medicine, New Haven, CT, USA*
9. *BioMag Laboratory, HUS Medical Imaging Center, Helsinki University Hospital, Finland*
10. *Department of Psychiatry, University of Helsinki and Helsinki University Hospital, Helsinki, Finland*

*These authors contributed equally.

Summary

Background Neurocognitive dysfunction—a predictor for poor outcomes and a driver of disability and costs—has remained a significant unmet need in individuals with major depressive disorder (MDD). As such, neurocognitive training could serve as a new therapeutic approach for alleviating symptoms and improving real-life functioning in MDD. Action video games are effective up-regulators of brain plasticity and could thus provide an ideal medium for targeted and personalized neurocognitive training in digital therapeutics for MDD. Efficacy and functional outcomes evidence for neurocognitive training has, however, remained inconclusive.

Methods In this randomized, double-blinded, comparator-controlled, superiority trial, MDD patients in Finland were randomly assigned (1:1:1) to receive a 12-week game-based intervention with personalized closed-loop neurocognitive training (MEL-T01, “Meliora”), an essentially identical active comparator (MEL-S01, “Sham”) with reduced training load, or treatment-as-usual alone. Both Meliora and Sham were provided as an adjunct to treatment-as-usual. The primary outcomes, superiority of Meliora over Sham and TAU, and superiority of Sham over TAU, were assessed using robust linear-mixed-model statistics of the change in depression symptoms measured with the Patient

Health Questionnaire-9 (PHQ-9) at 4-week intervals throughout the intervention. Lived experience experts were integral to all stages of this research. The trial was registered with ClinicalTrials.gov: NCT05426265.

Findings Between June 28, 2022 and August 14, 2024, 1,384 patients were screened for eligibility and 1,001 enrolled for the trial to receive either Meliora (337), Sham (347), or TAU alone (317) for 12 weeks. A per-protocol completer analysis included 483 patients of whom 315 (65·2%) identified as woman, 121 (25·0%) as man, and 47 (9·7%) as other. The mean age was 33·8 years (SD 9·5). For the primary endpoint (PHQ-9), Meliora was superior to Sham (adjusted mean difference c –0.513 (90% CI –1·007 to –0·018), Cohen's d –0·19, p=0·045), and both Meliora (c –1·138 (–1·542 to –0·735), d –0·43, $p=2{\cdot}2{\cdot}10^{-6}$) and Sham (c –0·626 (–1·034 to –0·218), d –0·24, p=0·006) were superior to treatment-as-usual alone. The superiority of Meliora over Sham (c –0·354 (–0·575 to –0·133), d –0·103, p=0·004) was further corroborated with an exploratory intention-to-treat analysis with full device-arm cohort of 684 patients. Adverse events, monitored with four channels, were reported by 148 of 1,001 patients (14·8%) with frustration (87 patients, 8·7%) being the most common. No serious adverse events were reported.

Interpretation These findings demonstrate the therapeutic potential of neurocognitive training in immersive game-based digital therapeutics, opening new avenues for scalable, personalized, and cost-effective depression treatments that improve functional outcomes.

Funding Business Finland, Research Council of Finland, Ella and Georg Ehrnrooth Foundation, Jane and Aatos Erkko Foundation, Sigrid Juselius Foundation, Aalto University, Kansaneläkelaitos Research Funding.

Introduction

A leading cause of disability globally, major depressive disorder (MDD) is a highly heterogeneous mental health condition commonly associated with symptoms such as persistent low mood, anhedonia, rumination, thought distortions, and negative bias. Neurocognitive dysfunction represents another category of central, yet often overlooked symptoms in MDD,[1] which includes impairments in executive functions such as cognitive control, attention, working memory, and processing speed.[2] These impairments play a central role in MDD-related disability, limit the efficacy of psychotherapies, and may persist even in remission from core mood symptoms.[2-6] Yet, conventional treatments such as antidepressants[7] and psychotherapy[8,9] primarily target core mood symptoms of MDD instead of neurocognitive dysfunction.[10] As such, neurocognitive functions such as executive control have remained an under-addressed domain in MDD[6] and an important untapped therapeutic target for improving broader functional outcomes.

"Executive Functions Training" (EFT) is an emerging approach to concurrently ameliorate both MDD symptoms and the neurocognitive dysfunction.[11,12] Supported by a rich literature on the neuroscience of cognitive control and executive functions, EFT addresses the MDD pathophysiology *per se* through improving the functioning of the fronto-parietal neural circuits underlying executive functions and induction of long-term neuroplasticity.[13] Subsequently, these improvements can further alleviate core mood and cognitive MDD symptoms through strengthened control over rumination, thought distortions, and negative biases, and restored ability for goal-directed behavior.[14] Closed-loop EFT can further elevate the effectiveness of such training by continuous measurement coupled with adjustment of the treatment to individual executive function levels that also dynamically change over time. This yields a standardized treatment that is not achievable with a human therapist and the measurement data further provide cognitive biomarkers and objective high-resolution measures of treatment efficacy. However, while initial scientific evidence supports the potential of EFT in MDD treatments,[11,12] a large fraction of the contributing studies may be insufficiently powered. Moreover, the lack of carefully designed control conditions[15] limits the attribution of the observed efficacy outcomes to the specific training approach as opposed to other differences between the investigational and control interventions such as activity type, usage time,[16] engagement,[17] and immersion,[18] and positive expectations.[19]

Action video games offer a powerful medium for neurocognitive training, linking endogenously motivating activity, sustained immersion, and effective drivers of brain plasticity.[20] In fact, even the playing of commercial action video games leads to a range of measurable training effects[16] despite being fully designed for entertainment and not usable "as is" for therapeutics. Building on this medium, neurocognitive training delivered in a video-game form has been shown to significantly improve cognitive functioning in healthy adults[21] and to improve cognitive functioning and clinical symptoms in attention-deficit hyperactivity disorder[22] as well as cognitive functioning in MDD. [23,24] However, both the specificity of comparator devices and the translation of cognitive gains into functional improvements and clinically meaningful reduction of symptoms have remained key objectives for further studies.[25]

Here, we assessed the therapeutic effects attributable to EFT in a video-game-based intervention for MDD. We developed an investigational device (MEL-T01, "Meliora") that appeared as an immersive action video game to the patient, but in the game mechanics incorporated closed-loop, personalized broad-spectrum EFT as the principal therapeutic mechanism-of-action. As putative auxiliary mechanisms of action, the device delivered a CBT-themed therapeutic game narrative and adaptive in-game activation that was comparable to behavioral activation in CBT (appendix). As an active comparator (MEL-S01, "Sham"), we used the same device, with identical personalized in-game activation and narrative, but with reduced EFT functionality (Suppl. Table 1, appendix). A rigorous comparator enabled this study to dissect the efficacy specifically attributable to EFT, thus addressing the putative limitations in earlier research. Both interventions were provided adjunct to treatment-as-

usual (TAU) in a randomized controlled trial (RCT) registered with ClinicalTrials.gov (NCT05426265). Using depression symptoms measured with the self-report Patient Health Questionnaire (PHQ-9), we asked whether Meliora was superior to Sham in alleviating depression symptoms, and whether both Meliora and Sham were superior to TAU alone.

Research in context

Evidence before this study

Cognitive dysfunction has remained a significant unmet need and a driver of disability and poor functional outcomes in major depressive disorder (MDD). We searched PubMed (last search Aug 14, 2025) for randomized trials on cognitive or neurocognitive training for MDD using terms covering neurocognitive training, executive functions training, cognitive control training, computerized cognitive training, and cognitive remediation (search string in appendix). We excluded neuromodulation-augmented interventions and conditions, such as dementia, where depression was a co-morbidity. Prior work shows small-to-moderate symptomatic and functional benefits from neurocognitive training in depression, with heterogeneous control conditions and modest sample sizes with a median of 79 and a maximum of 626 patients in intention-to-treat (ITT) cohorts (median 43, maximum 65 in per-protocol cohorts). In particular, rigorous comparators have remained scarce, and a vast majority of earlier studies are based on non-device control conditions and comparator devices with limited similarity in usage, immersion, user experience, and/or activity type. One study assessed positive imagery-based cognitive bias modification in 150 MDD individuals with a rigorous control condition and reported no difference in treatment efficacy between the investigational and comparator devices. Two studies based on a total of 222 individuals with depressive symptoms showed that attention-bias modification significantly reduces depression symptoms against a rigorous comparator condition. In line with these findings, recent meta-analyses highlight the variability in research design and outcomes and call for larger, higher-quality RCTs elucidating also functional outcomes. Previous research did not include RCT evidence linking the neurocognitive training outcomes with alleviation of symptoms or disability.

Added value of this study

To our knowledge, the present trial with 684 MDD patients in an ITT cohort and 483 in a per-protocol cohort is the largest randomized comparator-controlled trial testing the efficacy of personalized, closed-loop neurocognitive training as an adjunct to treatment-as-usual for adult MDD. We developed an action video game where closed-loop neurocognitive executive-functions training and therapeutic content were embedded into game mechanics and delivered in a manner that was immersive and engaging for the patient. To achieve a rigorous control condition, we derived an active comparator from the investigational device, which differed exclusively by reduced neurocognitive training load.

The investigational and comparator devices yielded indistinguishable adherence, total use time, and experienced immersion. A rigorous comparator strengthens the causal attribution of the observed symptoms and functional outcomes specifically to executive-function training rather than to effects of the video-game play, engagement, and intervention experience. Beyond demonstrating superiority over both the active comparator and treatment-as-usual in reducing depressive symptoms at the primary endpoint, the present study provides first evidence that individual neurocognitive training gains are associated with improvements in both depressive symptoms and functional impairment, directly linking the training effects with clinically meaningful outcomes.

Implications of all the available evidence

The present data, together with the earlier smaller and methodologically heterogeneous studies, indicate that personalized neurocognitive executive-functions training can yield clinically meaningful reduction in depression symptoms and concurrently alleviate functional disability in adults with MDD. Digital form of the intervention enables scalable and cost-efficient delivery. The derivative, immersion-matched comparator sets a methodological benchmark for future digital-therapeutic trials to disentangle outcomes attributable to specific training mechanisms from nonspecific effects.

Methods

Study design and participants

Meliora RCT was a remote, randomized, double-blinded, comparator-controlled, cross-over, add-on, three-arm clinical device trial with 18–65-year-old adults with an interview-confirmed MDD diagnosis. The study was approved from the Helsinki University Hospital (HUS) Regional Committee on Medical Research Ethics (HUS/3042/2021) and the Finnish Medicines Agency (FIMEA/2022/002976) and conducted in compliance with the Declaration of Helsinki. The study was pre-registered on ClinicalTrials.gov (NCT05426265). All patients gave informed written consent prior to participation.

The study patients were recruited in collaboration with Finnish specialized and occupational healthcare partners (Helsinki University Hospital, Turku University Hospital, and Mehiläinen Ltd.) with whom research permits were signed and where clinicians were encouraged to share information about the study to their patients. Patients were also recruited through social media (Facebook, Instagram, and Reddit), email campaigns, and physical posters. All recruitment channels guided the interested patients to the study website, where they digitally signed an informed consent form that sent the data to Helsinki University servers. Power analysis was conducted before the study commencement, assuming an effect size with Cohen's d 0·40, a significance level of α = 0·0167 (adjusted using the maximal Holm-

Bonferroni correction) and power of 0·80. This analysis indicated a minimum sample size of 400 patients for the primary endpoint.

Inclusion criteria were: age 18–65 years (see Table 1 for full details); diagnosis of MDD according to the Mini International Neuropsychiatric Interview (MINI) 6.0.0 module A [55] based on DSM-IV criteria; ongoing contact with mental health services; normal or corrected-to-normal vision (e.g., glasses or contact lenses); access to a Windows computer (manufactured in the 2010s or later) with internet and a mouse; and a valid email address and phone number for study communication. Exclusion criteria were: diagnosis of a psychotic or neurological disorder (e.g., epilepsy, brain injury; migraine was not exclusionary); active suicidality, assessed when necessary using the MINI module B (score ≥17); gaming addiction, evaluated when necessary using the Problem Gambling Severity Index (PGSI)[26] (score ≥8) and the Gaming Addiction Scale (GAS-7)[27] (≥4 responses of "sometimes" or more frequent); inability to provide informed consent; pregnancy or nursing; and status as an inmate or forensic patient. 232 patients did not meet these criteria and 147 patients withdrew voluntarily before participating, resulting in 1,001 patients enrolled in the study (see Figure 1, Table 1). After study enrollment, four patients asked for their data to be removed.

Randomization and masking

Between 28, June 2022 and 14, August 2024, 1,384 patients signed up for the study, who were then automatically randomized by the backend server in blocks of six to one of the three intervention arms (MEL-T01 ("Meliora"), MEL-S01 ("Sham"), and TAU) with a 1:1:1 ratio (see Figure 1), and evaluated for eligibility. A prespecified interim analysis was performed in August 2023, during which an interim cohort was unblinded to the trial statistician for assessment of predefined early-stopping criteria, including safety (appendix). Blinding was maintained for all other investigators, study personnel, and participants until trial completion. Recruitment was terminated when the minimum number of per-protocol completers for each arm, considering the dropout rate, was estimated to be accepted.

Procedures

After enrollment, the patients in Meliora and Sham arms automatically received an email link with instructions to install the intervention software on their personal computer using a digital distribution service, Steam (Valve Ltd.), to be used at their home. The patients were instructed to use the intervention for 48 hours, and at least 24 hours in total during the 12-week intervention period at their convenience. Usage was limited to a daily maximum of 90 min to safeguard against excessive use.

All patients were automatically sent an email after 4 (T1), 8 (T2), 12 (T3), and 24 (T4) weeks asking them to respond to a battery of symptom scales within 2 weeks of receiving the email. During this time window, the intervention main screen instructed the patients to answer the questionnaires before they could continue accessing the intervention content. After responding to the T3-questionnaire, the

patients in Meliora and Sham arms could no longer use the intervention and entered a follow-up period. Patients in the TAU arm were first monitored for a 12-week period, and at T3, the TAU patients received an email link with instructions to install the Meliora or Sham software and began the intervention period.

The clinical teams provided reactive support to patients via telephone and email during business hours. Aligned with the Finnish Ministry of Social affairs decree, all patients were reimbursed with 50 € if they met the 24 h intervention use goal and 120 € if they met the 48-h goal, if they also completed the symptom scales at T1, T2 and T3.

The investigational device, Meliora, was an action-video-game-based digital intervention with mechanisms of action including: 1. personalized, closed-loop EFT, 2. the intervention being behaviorally activating, and 3. having a CBT-inspired narrative (appendix). The active comparator, Sham, was otherwise identical to Meliora but with EFT elements disabled, leading to reduced neurocognitive training load while fully maintaining the behaviorally-activating and therapeutic-narrative driven elements as well as the entertainment-game-like presentation and intervention experience. A comparison between Meliora and Sham is provided in Supplementary Table 1.

Outcomes

The primary endpoint was the change in depressive symptom severity, as measured with Finnish translation of the PHQ-9 symptom scale,[28] where the change from baseline was calculated for each measurement point during the 12-week intervention. The secondary endpoints were Finnish measurements of depressive symptom severity with Self-Report Quick Inventory of Depressive Symptomatology[29] (QIDS-SR16), anxiety severity with Generalized Anxiety Disorder Questionnaire[30] (GAD-7), rumination with Short-Version Ruminative Response Scale (RRS-SV,[31]), disability with Sheehan Disability Scale[32] (SDS), quality-of-life with The World Health Organization-Five Well-Being Index[33] (WHO-5) and positive emotions / anhedonia with Positive Valence System Scale short form[31] (PVSS-SF). For neurocognitive measurements of performance in domains of planning, working memory, and short-/long-term memory, see appendix.

Statistical analysis

All primary and secondary hypotheses were pre-specified as superiority tests: Meliora vs TAU, Sham vs TAU, and Meliora vs Sham in reducing symptom severity or enhancing well-being. The per-protocol (PP) cohort included patients with PHQ-9 assessments at T2 and/or T3 and ≥ 24 h cumulative intervention usage. For patients who accumulated ≥24 h of use before cessation, the first symptom assessment after cessation was used. The primary outcome measure was change in symptom scores from baseline. Mean changes were plotted with 5th and 95th CIs from 10,000 bootstrap samples, drawn separately for each timepoint and study arm.

Missing data were imputed using Multiple Imputation by Chained Equations (MICE) with *miceforest* (random forest-based implementation; Supplementary Methods). As robust LMMs cannot be pooled with Rubin's rules, a single imputed dataset was selected based on MCSE convergence diagnostics.

Primary analysis used robust Linear Mixed Models (rLMMs) via the *robustlmm* R package *(rlmer*, DASVar estimator). Standard LMM diagnostics (Suppl. Table 6) indicated violations of normality and/or homoscedasticity, requiring robust estimation to meet protocol requirements. Fixed effects included timepoint, study arm, baseline symptom severity, income, life status, gender, age, and education (per regulatory recommendations[67,68]). Random intercepts were specified at the patient level.

Variance inflation factors (VIF) for covariates were <1·99, below the VIF < 5[34] threshold, with cross-correlations r<0·55, supporting inclusion of all covariates. Standardized effect sizes (Cohen's d) were calculated as the fixed-effect coefficient divided by the square root of the residual variance, following mixed-model conventions.

Holm-Bonferroni correction was applied separately to primary and secondary outcome families. Degrees of freedom (DoF) were approximated with Satterthwaite-like methods for LMMs and used for *p*-value calculation in robust Wald-type tests. CIs for fixed-effect estimates were computed from model-based SEs and the 5th and 95th percentiles of the Wald distribution. Differences in continuous demographic and clinical variables were tested with one-way ANOVA; categorical variables with independent categories used χ^2 tests. For categorical variables violating independence (e.g., medication, diagnoses, treatment contacts), p-values were obtained via a 10,000-permutation test, shuffling study arm labels and computing variance in proportions to generate the null distribution. The same statistical approach was used for all the secondary outcomes. Sensitivity analysis (complete case) for the key contrast (Meliora vs. Sham) confirmed robustness to the imputation procedure (c –0·554, d –0·218. p=0·037).

Intention-to-treat (ITT) and modified intention-to-treat (mITT-1 and mITT-2) analyses. The ITT analysis included all patients randomized to Meliora or Sham, regardless of intervention adherence (appendix). The first modified intention-to-treat (mITT-1) analysis included patients who (i) received active antidepressant treatment as part of treatment as usual and (ii) were aged 22–65 years. The second modified intention-to-treat (mITT-2) analysis applied the mITT-1 criteria plus at least moderately severe depressive symptoms at baseline (PHQ-9 ≥15). All other analytic procedures were identical to those described in "Statistical analysis".

Usage and immersion analysis. The relationship between subjective immersion, device usage, and treatment response was examined using regression on PHQ-9 scores at the 12-week (T3) measurement point, pooling data from the Meliora and Sham arms. Analyses followed the imputation procedures described in "Statistical analysis". Predictors were total intervention usage and average self-reported

immersion (IEQ), both standardized prior to analysis. Demographic covariates and baseline symptom severity were included as in the primary endpoint model. An interaction term between usage and immersion tested whether the effect of usage on symptom reduction varied by immersion level. Due to violations of normality and homoscedasticity, models were estimated using robust linear regression with Huber weighting. Standardized effect sizes (Cohen's *d*′) were calculated as fixed-effect coefficients divided by the square root of the residual variance.

Immersion-based split-cohort analysis. All accepted participants in the Meliora (n=337) and Sham (n=347) arms were split at the median average immersion score (IEQ) into High-IEQ and Low-IEQ subgroups. Comparisons were made between High-IEQ and Low-IEQ subgroups within each arm, and the primary endpoint analysis was repeated within the High-IEQ cohorts of both arms. Analyses otherwise followed the procedures described in "Statistical analysis".

Cognitive metrics analyses. Cognitive performance data were extracted aggregated per patient and game round. For time-based (non-bounded) metrics, outliers were excluded at the round level if they exceeded 3 SD from the group mean per timepoint. Measurements from rounds preceding the unlock of specific game features were excluded as missing values, as were the first 10 rounds after the tutorial to control for initial learning effects.

To control differences in general gaming proficiency and motor performance, actions-per-minute (APM) served as a proxy and was regressed out of cognitive measurements. For metrics used to drive difficulty adaptation (e.g., planning efficiency), the adaptation level itself was also regressed out to separate measurement variance from adaptation. Baseline depression severity (PHQ-9) was controlled for. After preprocessing, robust linear models (RLMs) were fit individually for each patient and cognitive variable to estimate change over time (slope) across 200 game rounds for round-based measurements, and across 18–30 distinct gaming days for day-based measurements. Slopes were tested against zero using one-sample *t*-tests.

The relationship between cognitive change and symptom improvement was assessed in the per-protocol (PP) cohort. Missing PHQ-9 values were handled as described in "Statistical analysis". The 12-week PHQ-9 score was used as the clinical outcome, and Spearman rank correlations were computed between the individual cognitive- and symptom slopes. To obtain an aggregate inferential measure across cognitive domains, p-values from the symptom–cognition correlations were combined using the Aggregated Cauchy Association Test[35] (ACAT), which provides robust p-value combination while accounting for potential dependence structures between tests. Each p-value was weighted and pooled, yielding a global test statistic and combined p-value summarizing the overall association between cognitive trends and symptom change. Descriptions of the tasks are provided in appendix.

Cost-effectiveness analysis

We conducted an exploratory cost-effectiveness analysis (appendix) comparing Meliora to Sham over 52 weeks, from joint healthcare-payer and societal perspectives, using the mITT-2 cohort. All costs are reported in 2024 USD. Intervention costs were fixed at $300 per arm. Direct healthcare costs were estimated from PHQ-9 using a convex quadratic mapping calibrated to published anchors[36]. Productivity losses were estimated from the Sheehan Disability Scale (SDS) absenteeism and presenteeism items. Health utilities were mapped from PHQ-9 and GAD-7 to EQ-5D-5L values (US tariffs)[76,] and QALYs were computed by trapezoidal integration and extrapolated under four post-treatment scenarios (stable, conservative, pessimistic, optimistic). Total costs included intervention, productivity, and direct-care components. Group differences in cost and QALYs were baseline-adjusted (ANCOVA-style). Net monetary benefit (NMB), incremental cost-effectiveness ratios (ICERs), and return on investment (ROI) were calculated. Cost-effectiveness acceptability curves (CEACs) were produced across a willingness-to-pay range of $0–150,000 per QALY.

Role of the funding source

The funding bodies did not have any role in study design, data collection and analysis, data interpretation, or writing of the manuscript.

Results

Patient disposition

Between June 28, 2022 and August 14, 2024, 1,384 patients were screened for eligibility for the randomized, double-blinded, comparator-controlled trial and randomized in a 1:1:1 ratio to receive either the investigational device (MEL-T01, “Meliora”), a highly similar active comparator (MEL-S01, “Sham”), or TAU alone for 12 weeks (Fig. 1). Both Meliora and Sham were provided as an adjunct to TAU. 1,001 patients were enrolled for the study. Following the intervention period, patients in the Meliora and Sham arms entered a 12-week follow-up, while those in the TAU arm received Meliora or Sham for 12 weeks at a 1:1 randomization ratio (Fig. 1). Symptom assessments were acquired five times during the 24-week study period (Suppl. Fig. 1). The patients exhibited a large diversity of co-morbidities and TAU treatments (Suppl. Fig. 2). The treatment arms did not differ significantly in terms of any demographic or clinical background variables (Table 1, Suppl. Tables 2 and 3).

Treatment completion and usage

The patients were instructed to use the intervention for 48 h. To be included in the per-protocol data analysis (completer cohort), the patients were required to use the intervention for 24 h and fill the symptom questionnaires at 8- (T2) and/or 12-week (T3) time point. This criterion was met by 30.6%

(Meliora arm) and 28.2% (Sham arm) of patients with no significant difference (p=0·68). This reflects a likely real-world adherence as the compliance in the trial was not enforced with email, text-message, or phone-call reminder from study personnel. The average total usage times in the Meliora and Sham arms (Table 2) did not differ either among the patients included in the analysis (45·5 h (SD 16·0), 44·9 h (SD 17·6), respectively, p=0·80) or across the whole cohort (18·4 h (SD 21·6), 18·2 h (SD 21·2), respectively, p=0·94).

Immersion

Equal usage and adherence imply that the active and comparator devices had similar behavioral engagement. To evaluate whether the user experience in terms of immersion differed between the two devices, we used the Immersive Experience Questionnaire[18] (IEQ) that quantifies cognitive involvement, real-world dissociation, challenge-skills balance, emotional involvement, and experience of control and autonomy. "Immersion" thus refers to the psychological state of deep involvement in the digital experience, characterized by focused attention, reduced awareness of the external environment, and a sense of being absorbed in and carried away by the activity. Mean IEQ scores did not differ significantly between Meliora and Sham either in the completer cohort (125·4 (SD 25·0), 123·1 (SD 24·7), respectively, p=0·53) (Table 2) or in the whole cohort (118·9 (SD 27·1), 115·9 (SD 27·0), p=0·25). In line with equivalent usage, these findings show that Meliora and Sham offered an equally immersive intervention and game experience, confirming the validity of Sham as a rigorous comparator for isolating the effects specifically attributable to the executive functions training in Meliora.

Primary outcomes

Using robust-linear-mixture-model- (rLMM) based statistical inference and Holm-Bonferroni correction for multiple comparisons, we observed significant differences between the Meliora (99 patients), Sham (96), and TAU (288) in alleviation of depression symptoms (PHQ-9) in the per-protocol completer cohort of 483 patients. Meliora was superior to Sham (p=0·045) and TAU ($p=2{\cdot}2{\cdot}10^{-6}$), and Sham was superior to TAU (p=0·006) (Fig. 2, Table 3, adjusted mean differences given by the rLMM c coefficients are shown in Table). The study thus met its predefined primary endpoints. An exploratory follow-up analysis showed that these effects were sustained at the in-trial follow-up (T4, Suppl. Fig. 1) 12 weeks after the end of the intervention (T3) with significant further symptom reduction (T3 PHQ-9 mean score change: –3·05 (95% CI –3·90 to –2·21); T4 PHQ-9 mean score change: –3·9 (95% CI –4·80 to –3·02), p=0·019) and Sham (T3: –2·00 (–2·78, –1·25); T4: –3·04 (–3·82, –1·25), p=0·012).

The rLMM analysis controlled for baseline symptom severity, income, education, life status, gender,

and age as fixed effects. Baseline symptom severity showed a significant medium-sized effect (*c* –1·161, Cohen's d –0·440, $p=1{\cdot}9{\cdot}10^{-29}$) for the primary outcome, indicating that every standard deviation of increasing baseline severity was associated with a 1·161 points greater symptom reduction in PHQ-9. Education (c –0·211, d –0·08, p=0·039) and income (c –0·248, d –0·094, p=0·034) levels showed small effects suggesting that patients with higher education and income levels achieved greater reductions in depression symptoms. The effects of life status, gender, and age were not significant.

Secondary outcomes: symptom scales

Predefined secondary endpoints included both depressive and depression-related symptoms assessed with self-report symptom scales. In alleviation of depression symptoms, as measured with the Self-Report Quick Inventory of Depressive Symptomatology (QIDS-SR16), Meliora was superior to Sham (p=0.012), and both Meliora ($p=2{\cdot}0{\cdot}10^{-9}$) and Sham ($p=8{\cdot}4{\cdot}10^{-4}$) were superior to TAU, which corroborates the primary endpoint observations (see Table 3). At the level of raw *p* values, Meliora was also superior to Sham in alleviating rumination (Short-Version Ruminative Response Scale, RRS-SV). Both Meliora and Sham were superior to TAU in alleviating quality-of-life (The World Health Organization-Five Well-Being Index, WHO-5), and anhedonia (Positive Valence System Scale short form, PVSS-SF). Furthermore, Meliora was superior to TAU in alleviating anxiety (GAD-7), and disability (SDS) (see Table 3). The impact of game-based interventions thus extended beyond depressive symptoms to transdiagnostic features of MDD, including cognitive-affective and functional domains.

Secondary outcomes: device usage time and immersion

To investigate the effects of immersion and device usage on treatment response, we pooled data from all accepted patients in the Meliora and Sham arms and conducted a regression analysis, controlling for baseline symptom severity, age, gender, income, life status, and education.

Immersion had a significant direct effect on PHQ-9 symptom reduction at the 12-week measurement point (c –0·676, d –0·158, p=0·003), indicating that each standard-deviation increase in immersion was associated with a 0·676-point greater reduction in depressive symptoms. Device usage time alone, on the other hand, did not show a significant direct effect (p=0·696). However, we observed a significant interaction between usage and immersion (c –0·752, d –0·176, $p=2{\cdot}8{\cdot}10^{-4}$) suggesting that the effect of usage on symptom reduction was dependent on the level of immersion, which strongly suggests that greater engagement with the intervention improves outcomes only if the experience is immersive.

As a post-hoc quantification of the role of immersion, we performed a split-cohort re-analysis of the PHQ-9 outcomes by splitting the Meliora and Sham cohorts at median immersion. The high-immersion

split cohort revealed significantly better outcomes compared to the low-immersion split cohort in Meliora (c –1·200, d –0·445, p=0·024), and marginally better in Sham (c –0·755, d –0·339, p=0·082). The high-immersion split-cohorts also replicated the primary outcome observations (Suppl. Fig. 2).

Intention-to-treat analysis (ITT) for the Meliora-Sham comparison

To evaluate the robustness and generalizability of the per-protocol findings of the Meliora-Sham differences, we performed an exploratory intention-to-treat analysis including all 684 patients randomized into the Meliora (337) and Sham (347) arms regardless of device usage. Missing data were handled by using a multiple imputation process with chained equations (MICE, see Methods). Meliora was superior to Sham in alleviating the symptoms of depression as measured with PHQ-9 (c –0·354, d –0·103, p=0·004) as well as in alleviating anxiety symptoms (GAD-7, c –0·286, d –0·104, p=0·004), and rumination (RRS-SV, c –0·360, d –0·136, $p=2{\cdot}9{\cdot}10^{-4}$) (Table 4). These results confirm that the specific therapeutic effects of EFT in Meliora extend to the broader population beyond the completer cohort.

Modified intention-to-treat analyses (mITT)

To further assess generalizability, we conducted two additional exploratory modified-ITT (mITT) analyses reflecting clinically relevant end-user populations. The first (mITT-1), including all 329 Meliora- (159) and Sham-arm (170) patients aged 22–65 and receiving antidepressants as treatment-as-usual, showed that Meliora was superior to Sham in alleviating depression symptoms measured with both PHQ-9 (c –0·371, d –0·123, p=0·016) and QIDS-SR16 (c –0·416, d –0·191, $p=4{\cdot}6{\cdot}10^{-4}$) as well as in alleviating anxiety symptoms (GAD-7, c –0·755, d –0·303, $p=1{\cdot}1{\cdot}10^{-7}$), and rumination (RRS-SV, c –0·310, d –0·131, p=0·011) (see Table 4). The second (mITT-2) included all 190 patients of the mITT-1 cohort who had moderately-severe or severe depression (baseline PHQ-9 ≥ 15). mITT-2 showed that Meliora (91 patients) was again superior to Sham (99) in alleviating depression symptoms as measured with PHQ-9 (c –1·062, d –0·363, $p=8{\cdot}6{\cdot}10^{-4}$) and QIDS-SR16 (c –1·068, d –0·477, $p=1{\cdot}2{\cdot}10^{-9}$) as well as in alleviating anxiety symptoms (GAD-7, c –0·719, d –0·290, p=0·002), and well-being (WHO-5, c 0.564, d 0·284, p=0.006) (see Table 4).

These findings further support the generalizability of the per-protocol findings and the clinical utility of game-based EFT in real-world populations. Consistent replication across per-protocol, split, ITT, and mITT cohorts thus provides strong evidence for that the intervention is effective across a diverse patient population irrespective of adherence criteria, baseline symptom or demographic characteristics, or missingness patterns.

Evidence for neurocognitive training effects

The therapeutic rationale and mechanism-of-action for neurocognitive training posits that (1) engaging with the training device induces improvements in neurocognitive functioning, and that (2) these gains mediate the alleviation of mood/cognitive symptoms and improvement of functional outcomes.

An exploratory analysis showed that Meliora improved cognitive functioning in the trained domains of planning, memory, and cognitive control (see Methods; Supplementary Table 4). The greatest cognitive improvements were observed in planning speed (Cohen's d –0·85; –22·3%), working memory retrieval time (d –1·25; –15·9%), and processing speed (d –0·67; –25·7%). To test whether the cognitive gains were associated with alleviation of symptoms, we aggregated evidence across all cognitive domains using multiple dependent correlations with ACAT (Methods).

Providing validation for the present approach, we found a positive association of the cognitive training effects with alleviation of symptoms (p=0·005). *Post hoc* tests showed that the cognitive improvements were associated with improved efficacy and greater reductions in depressive symptoms (p=0·008 to 0·024), anxiety (p=0·006), and functional impairment (p=0·008 to 0·026) (Supplementary Table 4). In addition, we also used ACAT to separately test the relationship of aggregated neurocognitive improvements and the efficacy of the intervention in alleviating mood symptoms (PHQ-9) and disability (SDS). The neurocognitive improvements predicted both the improvements in the PHQ-9 (*p*=0·0017) and SDS (p=0·001), further corroborating the notion that the impact of neurocognitive training, as a mechanism-of-action in itself, may both reduce symptoms and improve functional outcomes.

Cost-effectiveness of neurocognitive training

Cost-effectiveness analysis was performed with the mITT-2 cohort representing the delivery of Meliora as an adjunct to antidepressants in adult patients with severe MDD. The analysis used Sham as the comparator intervention to isolate specifically the cost-effectiveness attributable to EFT in a digital intervention. Compared to Sham, Meliora yielded savings of –$818 (95% CI –$1,275 to –$377), p=0·0012) in healthcare costs and savings of –$2.702 (95% CI –$7,098 to –$1,755), p=0·159) in productivity costs. For productivity, Meliora significantly decreased both presenteeism (p=0·004, Suppl. Fig. 5A) and absenteeism (p=0·020, Suppl. Fig. 5B), while Sham did not (p=0·150 and 0·086), although the difference between Meliora and Sham was not significant (p=0·150 and 0·311).

Total cost savings were –$3,530 (95% CI –$8,018 to $1,023), p=0·1006). Meliora saved an average of 0·0349 (95% CI 0·0233 to 0·464), p<0·0001) QALYs over 12 months at a negative incremental-cost-effectiveness-ratio (ICER) of –$101,932. Cost-effectiveness acceptability showed that compared to Sham, Meliora had 96·8% chance of being cost-effective at the willingness-to-pay threshold of $50,000/QALY and 99·5% at $100,000/QALY with a net monetary benefit of $7,023 at willingness-to-pay of $100,000/QALY (Suppl. Fig. 5C, D). Thus, EFT delivered in a digital intervention as an

adjunct to treatment-as-usual may thus yield significant cost savings and health benefits in a cost-effective manner.

Safety

Adverse events (AEs) were reported by 148 of 1,001 patients (14·8%) using the Meliora or Sham device either during the intervention period or after the crossover during the monitoring period. Frustration (87 patients, 8·7%), stress (33 patients, 3·3%), anxiety (28 patients, 2·8%), nausea (21 patients, 2·1%), and headache (6 patients, 0·6%) were the most common AEs (Supplementary Table 5). Out of all patients provided with the Meliora and Sham devices, 15·7% (77/490) and 13·9% (71/511), respectively, reported AEs (see Supplementary Table 5), with no significant difference (p=0·52). No serious AEs were reported.

Discussion

This superiority trial evaluated a novel digital intervention for MDD, MEL-T01 ("Meliora"), which comprised closed-loop EFT as well as video-game-based activation and a CBT-inspired therapeutic narrative. To dissect the putative therapeutic effects specifically attributable to EFT, we developed a novel, rigorous comparator (MEL-S01, "Sham") that was otherwise identical to the investigational device but had reduced functionality for adaptive EFT (see Suppl. Table 1). Both Meliora and Sham were delivered over a 12-week intervention period as an adjunct to TAU. The two devices achieved indistinguishable usage time, adherence, and immersion. Meeting all primary endpoints of the trial, we found that Meliora was superior to both Sham and TAU alone, while also Sham was superior to TAU alone in alleviating depressive symptoms measured by the PHQ-9 symptom scale. These findings were replicated with the QIDS-SR16 symptom scale as a secondary endpoint, establishing that the intervention effectively targeted depression as a construct.

As a further confirmatory approach, an exploratory intention-to-treat analysis further showed that these per-protocol findings generalized to the complete study population. The pre-registered plan to include into analyses only patients with at least half of the targeted 48h intervention usage time (completer cohort) was motivated by meta-analytic evidence showing that cognitive training requires 30–50 h of training time to yield consistent effects[16]. However, the reproducibility of the primary-outcome findings in an intention-to-treat analysis with full cohort shows that strict adherence criteria were not crucial for the intervention outcome. In addition to depression symptoms, Meliora was superior to Sham also in alleviating rumination and anxiety suggesting that also the secondary findings generalize beyond the completer population. These convergent lines of evidence for superiority of Meliora over

Sham thus provide compelling evidence for EFT-focused neurocognitive training as a specific mechanism of action in digital therapeutics for MDD.

The usage of DMHIs varies widely and can often be low,[37] while higher usage is associated with greater mental health improvements.[38] We included intervention usage and immersion as secondary endpoints. Interestingly, the efficacy of Meliora and Sham was strongly associated with immersion so that each standard deviation increase in immersion yielded a 0·68 PHQ-9 point greater reduction in depressive symptoms. Usage time, and thus behavioral engagement, had no significant effect alone but exhibited an interaction with immersion, suggesting that higher usage improves outcomes but only when paired with an immersive experience. This result supports and extends the understanding of how several subjective factors contribute to the effectiveness of digital therapeutics.[39] Moreover, this finding emphasizes how essential high-quality implementation is for investigational interventions to generate both subjective enjoyment and continued usage, which interact reciprocally with the mechanisms of action to drive the clinical benefits.[40] Yet, the development of interventions that meet the consumer-grade expectations of patients is highly challenging in academic settings and dependent on co-operation with industry and entertainment professionals. Finally, as engagement and immersion factors are essential for efficacy, it is crucial to have strong control over them with well-matched comparators in clinical trials.

We benchmarked the efficacy of Meliora, as an unsupported adjunct DMHI, against efficacy values reported for the Finnish psychotherapy quality register study[41] with 1,844 patients. Short psychotherapy (up to 20 sessions) in primary care achieved a –3·00 PHQ-9 point change from a baseline of 9·08 (Cohen's d –0·65). In psychiatric specialty care, short psychotherapy achieved a change of –2·50 from a baseline of 10·01 (d –0·50) and long psychotherapy (up to 40 sessions annually for three years) a change of –2·61 points from a baseline of 12·73 (d –0·43). In the per-protocol cohort, we found Meliora to achieve a PHQ-9 score change of –3·05 from a baseline of 15·4 (d –0·60) and, in the full intention-to-treat cohort, score change of –3·52 from a baseline of 15·82 (d –0·68). Thus, the overall efficacy and effect sizes achieved with Meliora were well aligned with face-to-face psychotherapies provided in specialty healthcare, and in fact, the effect sizes for Meliora-arm patients exceeded the 95% confidence intervals of both short ($d_{2.5\%}$ –0·59) and long psychotherapies ($d_{2.5\%}$ –0·58) therein.[41] Considering digital depression interventions as another benchmark, a recent multi-verse meta-analysis shows that digital depression interventions achieve an effect size with Hedges' g of 0·25 when controlled against care as usual.[42] In the present study, we found Meliora to yield a Cohen's d –0·43 and Sham a Cohen's d –0·24 in comparisons against treatment as usual, further emphasizing the capacity of EFT as an effective mechanism of action in digital therapeutics. Finally, an analysis assessing specifically the cost-efficiency of EFT showed that adjunctive digital

neurocognitive training significantly improves health outcomes and does so with a high probability of being cost-effective under conservative willingness-to-pay thresholds. The negative incremental-cost-effectiveness-ratio highlighted that Meliora generated both QALY gains and cost savings when compared with Sham.

Cognitive dysfunction may persist even in remission from mood symptoms[2–5] and is not alleviated by most current treatments.[7,10] Recent findings do, however, show that pharmacotherapy may yield limited alleviation in different cognitive domains, such as in learning and memory with Cohen's d of 0·25 in a meta-analysis.[43] Vortioxetine has demonstrated cognition-enhancing effects in MDD with similar small-to-moderate effect sizes (corrected for baseline severity, MADRS) across domains such as processing speed (d 0·14 to 0·34), executive function (d 0·24 to 0·27), and memory (d 0·06 to 0·21).[44] The investigational device in the present study yielded substantially larger effects sizes (controlling for baseline severity, PHQ-9) for training effects in executive planning time (d 0·85), processing speed (d 0·67), short-term (d 1·21) and long-term memory retrieval (d 0·40), working memory retrieval (d 1·25), and set-switching (d 0·80). While these training effects are not directly comparable with pharmacotherapy-associated improvements in cognition, they highlight the clinical potential of EFT in novel digital interventions aiming to alleviate both depressive symptoms and the cognitive dysfunction to beget improved functional outcomes. As key evidence for this notion, and a central validation for the therapeutic potential of neurocognitive training, the present data showed that the improvements in cognitive performance were predictive of both depression symptom alleviation and reduction in functional impairment (Sheehan disability).

As a notable limitation, this study was carried out as a remote trial with internet-based acquisition of symptoms data and hence the study lacks clinician-rated symptom endpoints. Self-report symptom scales have, however, been found to provide a conservative estimate of treatment effects compared to clinician ratings[45] and thus are more likely to under- than overestimate the intervention effects here. Moreover, observations of objective cognitive performance gains and their positive association with the changes in self-reported depression symptoms and disability measures further consolidate the present findings.

In conclusion, these findings suggest that executive functions training is an efficacious and cost-effective therapeutic mechanism of action for digital mental health interventions for adult major depressive disorder. The present study further highlights the potency and potential of immersive video games as media for delivering multi-action therapeutic interventions.

Acknowledgements

This research was funded by a Business Finland grant no. 215471/Z/19/Z and the Future Makers funding from Ella and Georg Ehrnrooth Foundation and Jane and Aatos Erkko Foundation awarded to J.M.P. The research was further funded by Aalto University funding to J.M.P., by Sigrid Juselius Foundation grants to J.M.P. and S.P., and by Research Council of Finland (305814) and Kansaneläkelaitos Research Funding grants to J.M.P. The views expressed are those of the authors and not necessarily those of Aalto University or the funding bodies. We thank all the trial patients who invested their time and effort to participate as well as the clinical organizations who referred and supported them. We are grateful to Tony Simon for insightful comments on the manuscript. We also extend our sincere gratitude to many contributors to this project who are not co-authors of this paper. Atte Hanski, Niilo Säämänen, and Ringo Leiden contributed to the design of the first implementation, *Meliora 1*. Atte Hanski and Carlos de la Guardia contributed to the design of *Meliora 2* and *3*, where Jhoser Buitrago and Miha Rinne contributed to game art, Lari Unkari, Juha Huotari, Vladislav Myrov, and Tuomas Puoliväli to software, Ilkka von Boehm to audio, and Anna Lampinen, Birgitta Paranko, and Laura Schildt to clinical study coordination. Jukka Laakso contributed to the design and production of *Meliora 4* ("Meliora", the investigational device of the present study), Sami Lehtinen, Oskari Rautala, Karri Kangas, Ville Ojala, Venla Lymysalo, and Flatfish Games Ltd. to software, Alpo Oksaharju to art, Mikko Rautalahti to narrative, and Tommi Hartikainen, Ville Ojala, and Unrivaled Audio Ltd. to audio. Leena Kähäri and Santeri Lepistö contributed to clinical study coordination.

Author Contributions

J.M.P. conceived the study. J.M.P. and S.P. obtained the funding. J.M.P, J.J.J., L.L., M.M., E.I., and S.P. designed the study and developed the study protocol. J.M.P., M.M., H.R., and E.I were principal investigators. M.V., V.-R.B., P.P., and A.S. enrolled and supported patients. J.J.J., A.S., L.P., and J.K. designed software. J.J.J. performed the data and statistical analysis and prepared the illustrations. J.M.P. wrote the first draft of the paper. All authors edited further revisions of the draft, approved the final paper, and agree to be accountable for the work.

Competing interests

J.M.P., J.J.J., and L.L. are shareholders in Soihtu DTx Ltd, which develops digital therapeutics for major depressive disorder and holds intellectual property transferred from Aalto University with J.M.P., J.J.J., L.L., and S.P. as co-inventors. J.M.P. is a part-time and J.J.J., L.L., A.S., J.K., and L.P. are active employees at Soihtu DTx. L.L. is a member of the Board of Directors of Soihtu DTx Ltd.

Soihtu DTx Ltd. did not play a role in the design, conduct, data analysis, or funding of the study. The other authors declare no competing interests.

References

1. Kriesche D, Woll CFJ, Tschentscher N, Engel RR, Karch S. Neurocognitive deficits in depression: a systematic review of cognitive impairment in the acute and remitted state. Eur Arch Psychiatry Clin Neurosci 2023;273:1105–1128.

2. Rock PL, Roiser JP, Riedel WJ, Blackwell AD. Cognitive impairment in depression: a systematic review and meta-analysis. Psychol Med 2014;44:2029–2040.

3. Pan Z, Park C, Brietzke E, et al. Cognitive impairment in major depressive disorder. CNS Spectr 2019;24:22–29.

4. McIntyre RS, Cha DS, Soczynska JK, et al. The prevalence, measurement, and treatment of the cognitive dimension/domain in major depressive disorder. CNS Drugs 2015;29:577–589.

5. Bortolato B, Miskowiak KW, Köhler CA, et al. Cognitive remission: a novel objective for the treatment of major depression? BMC Med 2016;14:9.

6. McIntyre RS, Konarski JZ, Wilkins K, et al. Cognitive deficits and functional outcomes in major depressive disorder: determinants, substrates, and treatment interventions. Depress Anxiety 2013;30:515–527.

7. Ou A, Wu GWY, Kassel MT, et al. Cognitive function in physically healthy, unmedicated individuals with major depression: relationship with depressive symptoms and antidepressant response. J Affect Disord 2025;378:191–200.

8. Zuckerman H, Pan Z, Park C, et al. Recognition and treatment of cognitive dysfunction in major depressive disorder. Front Psychiatry 2018;9:655.

9. Porter RJ, Bourke C, Carter JD, et al. No change in neuropsychological dysfunction or emotional processing during treatment of major depression with cognitive–behaviour therapy or schema therapy. Psychol Med 2016;46:393–404.

10. Keefe RSE, McClintock SM, Roth RM, et al. Cognitive effects of pharmacotherapy for major depressive disorder. J Clin Psychiatry 2014;75:864–876.

11. Woolf C, Lampit A, Shahnawaz Z, Aarons GA, Malhi GS, Harvey PD. A systematic review and meta-analysis of cognitive training in adults with major depressive disorder. Neuropsychol Rev 2022;32:419–437.

12. Gefen E, Launder NH, Davey CG, et al. Computerized cognitive training in people with depression: a systematic review and meta-analysis of randomized clinical trials. medRxiv 2024 (preprint).

13. Koster EHW, Hoorelbeke K, Onraedt T, Owens M, Derakshan N. Cognitive control interventions for depression: a systematic review of findings from training studies. Clin Psychol Rev 2017;53:79–92.

14. Fiorillo A, Carpiniello B, De Giorgi A, et al. Assessment and management of cognitive and psychosocial dysfunctions in patients with major depressive disorder: a clinical review. Front Psychiatry 2018;9:493.

15. Hoorelbeke K, Van den Bergh N, De Raedt R, Wichers M, Koster EHW. Preventing recurrence of depression: long-term effects of a randomized controlled trial on cognitive control training for remitted depressed patients. Clin Psychol Sci 2021;9:615–633.

16. Bediou B, Adams DM, Mayer RE, Tipton E, Green CS, Bavelier D. Meta-analysis of action video game impact on perceptual, attentional, and cognitive skills. Psychol Bull 2018;144:77–110.

17. O'Sullivan S, van Berkel N, Kostakos V, et al. Understanding what drives long-term engagement in digital mental health interventions: secondary causal analysis of the relationship between social networking and therapy engagement. JMIR Ment Health 2023;10:e44812.

18. Jennett C, Cox AL, Cairns P, et al. Measuring and defining the experience of immersion in games. Int J Hum Comput Stud 2008;66:641–661.

19. Peciña M, Chen J, Karp JF, Dombrovski AY. Dynamic feedback between antidepressant placebo expectancies and mood. JAMA Psychiatry 2023;80:389–398.

20. Bavelier D, Green CS. Enhancing attentional control: lessons from action video games. Neuron 2019;104:147–163.

21. Anguera JA, Boccanfuso J, Rintoul JL, et al. Video game training enhances cognitive control in older adults. Nature 2013;501:97–101.

22. Kollins SH, DeLoss DJ, Cañadas E, et al. A novel digital intervention for actively reducing severity of paediatric ADHD (STARS-ADHD): a randomised controlled trial. Lancet Digit Health 2020;2:e168–e178.

23. Anguera JA, Gunning FM, Areán PA. Improving late life depression and cognitive control through the use of therapeutic video game technology: a proof-of-concept randomized trial. Depress Anxiety 2017;34:508–517.

24. Keefe RSE, Cañadas E, Farlow D, Etkin A. Digital intervention for cognitive deficits in major depression: a randomized controlled trial to assess efficacy and safety in adults. Am J Psychiatry 2022;179:482–489.

25. Harvey PD. Digital therapeutics to enhance cognition in major depression: how can we make the cognitive gains translate into functional improvements? Am J Psychiatry 2022;179:445–447.

26. Ferris J, Wynne H. The Canadian Problem Gambling Index: final report. Ottawa: Canadian Centre on Substance Abuse; 2001.

27. Lemmens JS, Valkenburg PM, Peter J. Development and validation of a game addiction scale for adolescents. Media Psychol 2009;12:77–95.

28. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16:606–613.

29. Rush AJ, Trivedi MH, Ibrahim HM, et al. The 16-item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol Psychiatry 2003;54:573–583.

30. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166:1092–1097.

31. Khazanov GK, Ruscio AM, Forbes CN. The Positive Valence Systems Scale: development and validation. Assessment 2020;27:1045–1069.

32. Sheehan DV, Harnett-Sheehan K, Raj BA. The measurement of disability. Int Clin Psychopharmacol 1996;11:89–95.

33. Topp CW, Østergaard SD, Søndergaard S, Bech P. The WHO-5 Well-Being Index: a systematic review of the literature. Psychother Psychosom 2015;84:167–176.

34. Tamura R, Kobayashi K, Hyoudou R, Matsui T, Goh M. Mixed integer quadratic optimization formulations for eliminating multicollinearity based on variance inflation factor. J Glob Optim 2019;73:431–446.

35. Liu Y, Chen S, Li Z, et al. ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. Am J Hum Genet 2019;104:410–421.

36. Popkov AA, Barrett TS, Shergill A, Donohue M, Anderson RJ, Karlin BE. Association between depression symptom severity and total cost of care: findings from a large, 2-year, claims-based, retrospective population health study. J Affect Disord 2025;368:41–47.

37. Fleming T, Bavin L, Lucassen M, et al. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. J Med Internet Res 2018;20:e199.

38. Gan DZQ, McGillivray L, Han J, Christensen H, Torok M. Effect of engagement with digital interventions on mental health outcomes: a systematic review and meta-analysis. Front Digit Health 2021;3:764079.

39. Graham AK, Kwasny MJ, Lattie EG, Greene CJ, Gupta NV, Reddy M, Mohr DC. Targeting subjective engagement in experimental therapeutics for digital mental health interventions. Internet Interv 2021;25:100403.

40. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017;7:254–267.

41. Saarni SE, Rikala S, Priha O, et al. Finnish Psychotherapy Quality Register: rationale, development, and baseline results. Nord J Psychiatry 2023;77:455–466.

42. Plessen CY, Panagiotopoulou OM, Tong L, Cuijpers P, Karyotaki E. Digital mental health interventions for the treatment of depression: a multiverse meta-analysis. J Affect Disord 2025;369:1031–1044.

43. Ainsworth NJ, Dursun S, Andrews L, et al. Cognitive outcomes after antidepressant pharmacotherapy for late-life depression: a systematic review and meta-analysis. Am J Psychiatry 2024;181:234–245.

44. Harrison JE, Lophaven S, Olsen CK. Which cognitive domains are improved by treatment with vortioxetine? Int J Neuropsychopharmacol 2016;19:pyw054.

45. Miguel C, de la Vega R, Belloch A, et al. Self-reports vs clinician ratings of efficacies of psychotherapies for depression: a meta-analysis of randomized trials. Epidemiol Psychiatr Sci 2025;34:e15.

Code availability

The custom code used in this study is not publicly available but will be shared upon reasonable request, subject to review on a case-by-case basis. Requests should be directed to the corresponding author. All code was provided to editors and peer reviewers during the review process, in accordance with journal policy.

Data availability

The datasets generated and analyzed during the current study are not publicly available due to patient privacy concerns set forth in the Ethical Committee approval, General Data Protection Regulation (GDPR) constraints, and stipulations outlined in the Clinical Investigation Plan (CIP). However, preprocessed data can be deidentified and may be made available upon reasonable request, subject to applicable data protection laws, approval by the Helsinki University Hospital (HUS) Regional Committee on Medical Research Ethics, and the execution of a data-sharing agreement. Requests should be directed to the corresponding author and must include a scientifically and methodologically sound research proposal.

Inclusion and Ethics

As this is a national study, local researchers have completed study design, study implementation, data ownership, intellectual property and authorship of publications. The research has been achieved with the two largest national hospital districts and a private health care provider. The study has been approved by a local ethics review committee. Participants of the study received treatment-as-usual throughout their participation and were not vulnerable to stigmatization, incrimination, discrimination. Risk assessment and management was comprehensively accounted for in the clinical investigation protocol and implemented throughout the study.

Figure Legends

Fig. 1: CONSORT diagram. CONSORT diagram of all patients assessed for eligibility, randomized into the MEL-T01, MEL-S01, and TAU arms, and followed up to 12 weeks. A total of 1,384 individuals were screened, of whom 1,001 were accepted. The per-protocol analyses included patients who completed T2 and/or T3 symptom assessments and met the ≥24-hour device-usage threshold.

Fig. 2: Primary and secondary depression symptom endpoints. a, Time course of the mean symptom change for the primary endpoint (PHQ-9) and **b,** secondary endpoint (QIDS-SR16) across the Meliora, Sham, and TAU arms. Scores reflect the mean change from baseline at each measurement point over the 12-week intervention period. Shaded areas represent the 5th and 95th confidence intervals from 10,000 bootstrap samples.

Table 1: Demographics and baseline clinical characteristics of patients included in the per protocol analysis.

Values represent mean (SD) or percentages, as appropriate. No significant differences were observed between the Meliora, Sham, and TAU groups at baseline across demographic or clinical characteristics. SSRI = Selective serotonin reuptake inhibitors. SNRI = Serotonin and norepinephrine reuptake inhibitors. TCA = Tricyclic antidepressant. Social therapy includes: rehabilitative work, family therapy, occupational therapy, art therapy, activities with a support person, mindfulness, and exercise. TMS = Transcranial magnetic stimulation. ECT = Electroconvulsive therapy.

Category	Variable	Meliora (n=99)	Sham (n=96)	TAU (n=288)	p
Age	Range	18–57	18–63	18–65	
	Mean, SD	33·2 (8·9)	35·2 (10·4)	33·5 (9·4)	0·26
Gender	Woman	65·7 %	64·6 %	65·3 %	
	Man	24·2 %	26·0 %	25·0 %	1·0
	Other	10·1 %	9·4 %	9·7 %	
Antidepressants	SSRI	25·2 %	22·9 %	20·8 %	
	SNRI	15·2 %	18·8 %	14·9 %	
	Vortioxetine	13·1 %	12·5 %	12·8 %	0·93
	TCA	2·0 %	1·0 %	2·1 %	
	None	23·2 %	27·1 %	28·1 %	
Treatments	Psychotherapy	32·3 %	30·2 %	36·1 %	
	Supportive talk therapy	61·6 %	54·2 %	54·9 %	
	Social therapy	52·5 %	44·8 %	51·7 %	0·56
	Light therapy	1·0 %	4·2 %	4·2 %	
	TMS	2·0 %	3·1 %	0·3 %	
	ECT	2·0 %	3·1 %	1·7 %	
Diagnoses	Mild depression	8·1 %	10·4 %	12·5 %	
	Moderate depression	76·8 %	68·8 %	70·1 %	
	Severe depression	30·3 %	38·5 %	34·4 %	0·29
	Anxiety	40·4 %	49·0 %	51·4 %	
	ADHD	13·1 %	10·4 %	17·7 %	
Symptom scales	PHQ-9	15·4 (4·4)	15·3 (4·5)	15·0 (4·8)	0·74
	QIDS-SR16	15·1 (3·4)	15·0 (3·3)	14·7 (3·6)	0·55
	GAD-7	10·8 (4·7)	10·5 (4·5)	10·9 (4·6)	0·74
	GAS-7	11·5 (3·2)	11·5 (3·6)	11·6 (3·5)	0·96
	RRS-SV	19·4 (4·0)	19·8 (3·5)	19·6 (4·4)	0·81
	SDS	18·9 (5·4)	18·7 (5·9)	18·1 (6·0)	0·48
	WHO-5	5·6 (3·1)	5·6 (3·3)	6·2 (3·5)	0·19
	PVSS-SF	100·2 (28·5)	106·6 (29·9)	106·5 (29·6)	0·17
	BEAQ	59·4 (9·0)	59·9 (10·9)	58·7 (10·7)	0·58
	PCL-5	36·8 (14·6)	38·0 (14·1)	37·4 (14·8)	0·87
	ASSIST	1·8 (1·5)	1·7 (1·6)	1·8 (1·6)	0·74

Table 2: Device usage and immersion in the Meliora and Sham arms.

Total device usage hours and self-reported immersion scores (IEQ) are shown for both the per-protocol analysis cohort and the full included sample. Values in parentheses indicate SD. No significant differences were observed between groups on either measure.

Category	Cohort	Meliora	Sham	p
Device usage hours				
	Completer cohort	45·5 (16·0) h	44·9 (17·6) h	0·80
	All accepted	18·4 (21·6) h	18·2 (21·2) h	0·94
Immersion (IEQ)				
	Completer cohort	125·4 (25·0)	123·1 (24·7)	0·53
	All accepted	118·9 (27·1)	115·9 (27·0)	0·25

Table 3: Primary (PHQ-9) and secondary endpoints.

The three hypotheses (Meliora is superior to Sham, Sham is superior to TAU, and Meliora is superior to TAU) were tested with robust LMM. For each symptom outcome, the estimated coefficient (c), associated p-value, and standardized effect size (d) are reported. Values of c in parentheses indicate 95% confidence intervals. Statistical significance was assessed at a per-protocol alpha level of p=0.05, with Holm–Bonferroni correction applied across outcome families. * indicates uncorrected significance; ** indicates significance after correction.

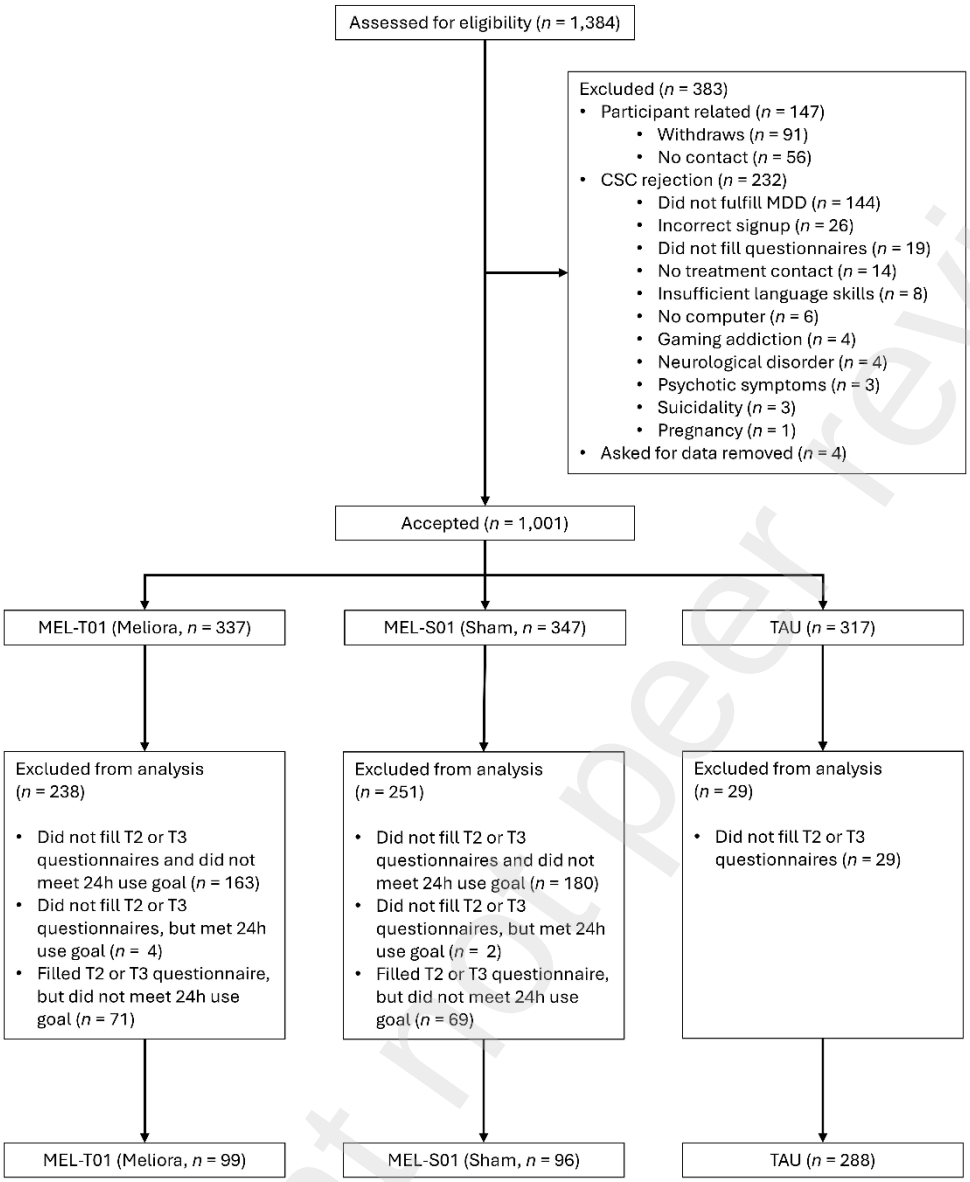
		Meliora superior to Sham			**Sham superior to TAU**			**Meliora superior to TAU**		
Symptom Scale	**Construct**	**c**	**p**	**d**	**c**	**p**	**d**	**c**	**p**	**d**
PHQ-9	Depression	–0·513 (–1·007, –0·018)	0·045**	–0·194	–0·626 (–1·034, –0·218)	0·006**	–0·237	–1·138 (–1·542, –0·735)	$2{\cdot}2{\cdot}10^{-6}$**	–0·431
QIDS-SR16	Depression	–0·524 (–0·901, –0·146)	0·012**	–0·247	–0·598 (–0·909, –0·287)	$8{\cdot}4{\cdot}10^{-4}$**	–0·282	-1·121 (–1·429, –0·814)	$2{\cdot}0{\cdot}10^{-9}$**	–0·529
GAD-7	Anxiety	–0·461 (–0·934, 0·012)	0·055	–0·188	–0·371 (–0·761, 0·019)	0·059	–0·151	–0·832 (–1·218, –0·447)	$2{\cdot}1{\cdot}10^{-4}$**	-0·400
SDS	Disability	–0·622 (–1·317, 0·073)	0·071	–0·162	–0·769 (–1·342, –0·196)	0·014**	–0·200	–1·391 (–1·958, –0·825)	$3{\cdot}1{\cdot}10^{-5}$**	–0·361
RRS-SV	Rumination	–0·403 (–0·799, –0·006)	0·048*	–0·182	0·233 (–0·094, 0·560)	0·879	0·106	–0·170 (–0·492, 0·153)	0·194	–0·077
WHO-5	Well-being	–0·055 (–0·404, 0·293)	0·603	–0·027	1·047 (0·760, 1·335)	$2{\cdot}1{\cdot}10^{-9}$**	0·502	0·992 (0·707, 1·276)	$8{\cdot}6{\cdot}10^{-9}$**	0·475
PVSS-SF	Positive valence	–0·429 (–3·238, 2·380)	0·599	–0·030	7·378 (5·070, 9·687)	$1{\cdot}1{\cdot}10^{-7}$**	0·514	6·950 (4·661, 9·238)	$4{\cdot}1{\cdot}10^{-7}$**	0·484

Table 4: Intention to treat (ITT) and modified ITT (mITT) analyses.

This table summarizes the effect of Meliora compared to Sham across the ITT, mITT-1, and mITT-2 cohorts. The mITT-1 cohort included patients aged ≥22 who received active antidepressant medication, while mITT-2 additionally required baseline PHQ-9 ≥ 15 (moderately severe or severe symptoms). For each symptom outcome, the estimated coefficient (c), associated p-value, and standardized effect size (d) are reported. All comparisons tested the hypothesis that Meliora was superior to Sham. * indicates $p<0.05$.

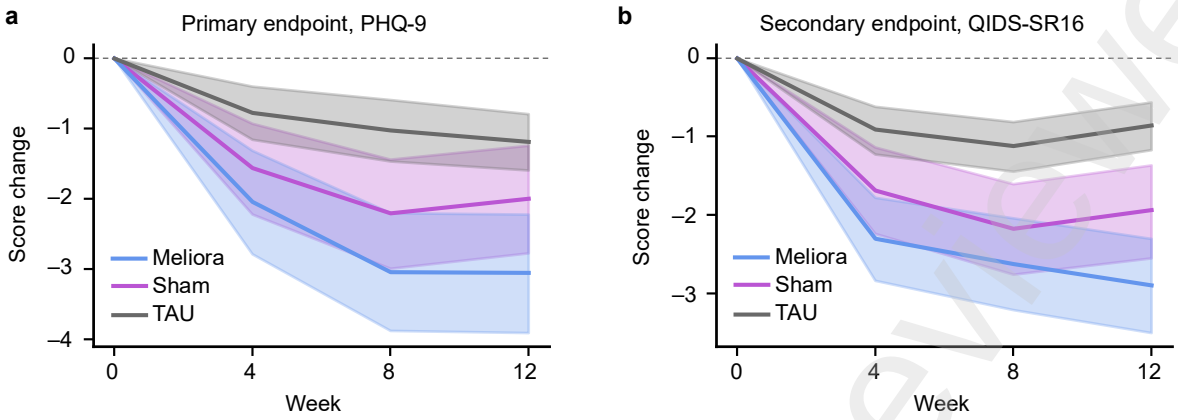
		ITT			mITT-1			mITT-2		
Symptom Scale	**Construct**	**c**	**p**	**d**	**c**	**p**	**d**	**c**	**p**	**d**
PHQ-9	Depression	·0·354 (–0·575, ·0·133)	0·004*	–0·103	–0·371 (–0·656, –0·087)	0·016*	–0·123	–1·062 (–1·611, –0·513)	$8·6·10^{-4}$*	–0·363
QIDS-SR16	Depression	·0·123 (–0·290, ·045)	0·114	–0·047	–0·416 (–0·621, –0·212)	$4·6·10^{-4}$*	–0·191	–1·068 (–1·348, –0·788)	$1·2·10^{-9}$*	–0·477
GAD-7	Anxiety	·0·286 (–0·465, ·0·108)	0·004*	–0·104	–0·755 (–0·989, –0·521)	$1·1·10^{-7}$*	–0·303	–0·719 (–1·123, –0·314)	0·002*	–0·290
SDS	Disability	·0·032 (–0·319, ·256)	0·428	–0·007	–0·282 (–0·652, 0·088)	0·105	–0·072	–0·066 (–0·503, 0·372)	0·402	–0·019
RRS-SV	Rumination	·0·360 (–0·531, ·0·189)	$2·9·10^{-4}$*	–0·136	–0·310 (–0·533, –0·087)	0·011*	–0·131	–0·273 (–0·587, 0·041)	0·078	–0·108
WHO-5	Well-being	·0·220 (–0·371, ·0·069)	0·992	–0·094	0·121 (–0·109, 0·351)	0·193	0·060	0·564 (0·199, 0·928)	0·006*	0·284
PVSS-SF	Positive valence	·1·109 (–2·516, ·298)	0·902	–0·070	–2·654 (–4·755, –0·554)	0·981	–0·175	–1·218 (–4·256, 1·820)	0·745	–0·081

Fig. 1: CONSORT diagram.



CONSORT diagram of all patients assessed for eligibility, randomized into the MEL-T01, MEL-S01, and TAU arms, and followed up to 12 weeks. A total of 1,384 individuals were screened, of whom 1,001 were accepted. The per-protocol analyses included patients who completed T2 and/or T3 symptom assessments and met the ≥ 24-hour device-usage threshold.

Fig. 2: Primary and secondary depression symptom endpoints.



a, Time course of the mean symptom change for the primary endpoint (PHQ-9) and **b,** secondary endpoint (QIDS-SR16) across the Meliora, Sham, and TAU arms. Scores reflect the mean change from baseline at each measurement point over the 12-week intervention period. Shaded areas represent the 5th and 95th confidence intervals from 10,000 bootstrap samples.

Supplementary Appendix

Supplementary Methods

Supplementary Figures and Tables

1. Search string for Research in Context

("major depressive disorder"[Title/Abstract]
OR “depression”[Title/Abstract]
OR “major depression”[Title/Abstract])

AND ("cognitive training"[Title/Abstract]
OR "computerized cognitive training"[Title/Abstract]
OR "computerized training"[Title/Abstract]
OR "cognitive remediation"[Title/Abstract]
OR "executive function training"[Title/Abstract]
OR "executive functions training"[Title/Abstract]
OR "cognitive control training"[Title/Abstract]
OR "attention training"[Title/Abstract]
OR "attention bias modification"[Title/Abstract]
OR "working memory training"[Title/Abstract]
OR "digital cognitive training"[Title/Abstract]
OR "game-based cognitive training"[Title/Abstract]
OR "computer program"[Title/Abstract]
OR ("digital intervention"[Title/Abstract]
AND ("cognitive deficit"[Title/Abstract]
OR "cognitive dysfunction"[Title/Abstract]
OR "cognitive impairment"[Title/Abstract])))

AND (“adult”[Title/Abstract]
OR “adults”[Title/Abstract]
OR “students”[Title/Abstract])

AND (“randomized”[Title/Abstract]
OR “randomised”[Title/Abstract]
OR “trial”[Title/Abstract]
OR "randomized controlled trial"[Publication Type])

NOT (“Mild cognitive impairment”[Title/Abstract]
OR "Dementia"[Title/Abstract]
OR "HIV"[Title/Abstract]
OR "ADHD"[Title/Abstract]
OR "Stroke"[Title/Abstract]
OR "Children"[Title/Abstract]
OR "tDCS"[Title/Abstract]
OR "open label"[Title/Abstract]
OR “study protocol”[Title/Abstract])

2. Meliora and Sham devices

The devices were developed at Aalto University, and the patients installed and operated them on their personal computers with a Windows operating system. As a delivery medium for the mechanisms of action, the intervention used a first-person action video game where the patient controlled an avatar in a complex 3-dimensional environment (Supplementary Figure 4) while interacting with environmental objects and adversarial encounters. The intervention included 28 levels, and progressing through the levels unlocked the narrative and new game features that allowed for more complex interactions. The patient was required to play a particular level for several rounds. Across the Meliora and Sham devices, the average game round duration was 9·87 min (SD 3.6), and the patients played an average of 182 (SD 72) rounds during the intervention with 135 out of 195 subjects (69 %) reaching the level 28. The adaptive EFT (see Cognitive tasks below) was implemented as game encounters and puzzles of which the difficulty level was adapted to patient performance between intervention rounds to achieve personalized, suitable difficulty level.

The rationale for using an action video game as the medium for delivering EFT, in addition to ensuring engagement and promoting immersion, was to offer experiences of pleasure and mastery and increase activation, which are core components in behavioral activation in CBT[46] and have been found effective treatment of depression.[47,47] As specific behavioral-activation-like digital intervention elements,[49] both devices (i) it had scheduled user milestones and positive feedback, (ii) its use was pleasurable, enjoyable, and satisfying, and iii) it enhanced the patient's sense of competence and agency through learning to solve in-game challenges.

Aligned with such behavioral activation, Meliora and Sham provided activity structuring and scheduling by encouraging the patient to use the intervention daily or several times a week. The intervention main menu visualized the intervention usage goals, achievements, and progression through the intervention levels. The CBT-inspired game narrative was implemented as a 28-part, character-driven, voice-acted and subtitled narrative that focused on identifying cognitive biases.

3. Missing data and imputation

Missing Data Mechanism. In the per-protocol cohort, missingness in the primary outcome (PHQ-9) was 20·2%. Logistic regression indicated associations with study arm ($p=2{\cdot}9{\cdot}10^{-54}$; OR=1·022) and age (p=0·025; OR=0·996), suggesting higher odds of missing data in the TAU arm and among younger participants. Model AUC was 0·71, indicating data were not Missing Completely at Random (MCAR) but plausibly Missing at Random (MAR) conditional on observed covariates. These findings support the MAR assumption required for multiple imputation.

Imputation Procedure. Missing data were imputed using Multiple Imputation by Chained Equations (MICE) with a random forest implementation (*miceforest*). The imputation model included age, gender, income, life status, baseline symptom severity, and treatment arm; continuous covariates were standardized.

Convergence was monitored via Monte Carlo Standard Error (MCSE) and the Fraction of Missing Information (FMI/m). After 160 iterations, MCSE stabilized at 0·075 (<0·10) and the FMI/m ratio at 0·006 (<0·02), indicating adequate convergence. Because robust LMMs cannot be validly pooled across multiply imputed datasets (due to data- and model-level variation), a single imputed dataset was selected based on convergence diagnostics, providing a representative dataset for analysis.

4. Intention-to-treat (ITT) and modified ITT Analyses

The ITT cohort included all randomized participants, regardless of adherence. Missingness in PHQ-9 was 44·4% and associated with education ($p=2{\cdot}9{\cdot}10^{-5}$; OR=0·998), baseline symptom severity (p=0·010; OR=1·007), and age (p=0·031; OR=0·994). Logistic regression AUC was 0·54, indicating weak predictive accuracy. These results support MAR conditional on observed covariates.

The same imputation procedure and diagnostics (MCSE, FMI/m, multicollinearity) were applied to ITT and both mITT cohorts: Modified ITT 1 (mITT-1) included all ITT patients receiving active antidepressant treatment as part of TAU, aged 22–65 years. Missingness (43·5%) was associated with education ($p=9{\cdot}6{\cdot}10^{-5}$; OR=0·996), life status (p=0·017; OR=1·002), baseline severity (p=0·014; OR=1·004), and age (p=0·042; OR=0·996); logistic regression AUC=0·55. Modified ITT 2 (mITT-2) included all mITT-1 patients with PHQ-9 ≥15 at baseline. Missingness (43·7%) was associated with education ($p=3{\cdot}4{\cdot}10^{-6}$; OR=0·830) and baseline severity ($p=4{\cdot}1\times10^{-5}$; OR=1·273); logistic regression AUC=0·61. Across ITT and mITT analyses, missingness was conditionally dependent on observed covariates, supporting the MAR assumption and use of multiple imputation.

5. Health economics analyses

Missing Data Assumptions. Missing data were assumed MAR, consistent with observed patterns (see "mITT-1 and mITT-2 analyses").

Productivity Costs. Productivity outcomes were derived from the Sheehan Disability Scale (SDS) so that Absenteeism was provided by SDS item 4 and Presenteeism by item 5 that was quantified as a half-day productivity loss. Productivity costs were estimated from the reported annual income in demographics data, scaled to 262 workdays/year and weighted by proportion of workdays lost.

Imputation and Resampling. Imputation was achieved with Bayesian Ridge regression with posterior sampling using 100 multiply imputed datasets. For bootstrapping, we generated 1,000 stratified bootstrap replications per dataset. For pooling, Rubin's rules were used for point estimates and SEs.

Health Utilities and QALYs. Health utilities were mapped from PHQ-9 and GAD-7 to EQ-5D-5L values using the ADVLDD mapping method (US tariffs; Franklin et al. 2023). QALYs were estimated by trapezoidal integration and extrapolated to 52 weeks.

Direct Healthcare Costs. Direct healthcare costs were estimated from PHQ-9 using a convex quadratic mapping calibrated to published anchor points.

Area-Under-the-Curve (AUC) Estimation. The trapezoidal AUC approach and scenario-based extrapolation were applied consistently to QALYs, productivity (days saved), and healthcare costs, ensuring coherence across domains.

Cost-Effectiveness Metrics. Total costs included intervention, productivity, and direct-care components. Group differences in cost and QALYs were adjusted for baseline values and income (ANCOVA-style regression). From these, we computed Net monetary benefit (NMB), Incremental cost-effectiveness ratios (ICERs), and Return on investment (ROI). Uncertainty was assessed by bootstrap resampling of imputed datasets. For sensitivity analyses, the cost-effectiveness acceptability curve (CEACs) was estimated across willingness-to-pay thresholds ($0–150,000 per QALY).

6. Cognitive tasks for behavioral performance assessment

The active intervention, Meliora, included a diverse set of adaptive cognitive tasks targeting planning, memory, and cognitive control. For the present study, we analyzed data from three tasks designed to be structurally comparable to established clinical paradigms or intended primarily as measurement rather than adaptive EFT.

Planning task (Tower of London–like). Patients were shown an initial and a goal configuration of objects arranged across three vertical poles, with variation in the number and arrangement of objects. They were instructed to plan and then execute the minimum sequence of moves to match the goal state. Progression was self-paced, and difficulty adapted dynamically based on prior performance and the optimal number of moves. Behavioral measures were: planning time (latency before the first move), execution time (time from first to final move, normalized by number of moves), and planning efficiency (ratio of actual to optimal moves).

Working-memory task. Patients completed a working memory task embedded in the game environment, with visual parameters determined by device-specific display properties. In each trial, three visually distinct geometric objects were presented without time limits for encoding. After a short

in-game navigation delay, patients selected the correct composite object (integrating features from all three stimuli) from three alternatives. Behavioral measures were: encoding duration, retrieval time, and accuracy.

Layered recall task (memory and cognitive control). This task probed short- and long-term memory with rule-based control demands, conceptually similar to the Wisconsin Card Sorting Test. It was administered across three consecutive gaming days in fixed blocks. On Day 1, patients identified the correct symbol among four by trial and error, progressing through three hierarchical layers; incorrect responses triggered a reset to the initial layer. On Days 2 and 3, the identical symbol set was used, requiring retrieval of the prior solution. Behavioral measures were: control errors (repeated selection of symbols previously identified as incorrect), retrieval times, and accuracies.

7. Interim analysis

A protocol-defined, conditional interim analysis was conducted after 23rd of March 2023 when each study arm Meliora, Sham, and TAU had at least 33 completers (n=47, n=40, n=120, respectively). The purpose of the interim analysis was to evaluate whether early study discontinuation conditions were met while maintaining the overall Type I error rate (α=0·05).

The "Clearly Beneficial" discontinuation condition assessed whether all primary hypotheses were achieved using a reduced alpha level of 0·025, with the remaining alpha (0·025) reserved for confirmatory testing. Statistical methods were consistent with those described in the "Statistical analysis" section. None of the primary hypotheses remained significant after Holm-Bonferroni correction (uncorrected p-values: Meliora vs. Sham, p=0·127; Meliora vs. TAU, p=0·029; Sham vs. TAU, p=0·323). As not all three comparisons were significant, this early stopping condition was not met.

The "Clearly Harmful" discontinuation condition assessed whether either investigational device (Meliora or Sham) demonstrated significantly less reduction in depressive symptoms compared to TAU with an alpha level of 0·05. Analyses followed the same statistical methods described in "Statistical Analyses." Neither Meliora (p=0·989) nor Sham (p=0·904) showed evidence of being clearly harmful, and this discontinuation condition was not met.

The "Futile" discontinuation condition qualitatively assessed patient recruitment, dropout, and adherence. None of the criteria were fulfilled: 1) We did not observe considerable difficulty in recruiting patients, as demonstrated by 638 recruited subjects in 10 months, 2) the dropout rate of 66·8% was within the expected range based on adherence criteria, and 3) patients who met the inclusion criteria engaged with the intervention for 46·6 ± 16·3 hours (Meliora) and 47·5 ± 14·7 hours (Sham) (mean ± standard deviation). Thus, the "Futile" discontinuation condition was not met.

Since no stopping decisions were made and no conclusions were drawn from the interim analysis, the full α=0·05 remained available for the final analysis, as no alpha was spent at this stage.

8. Additional references for Appendix

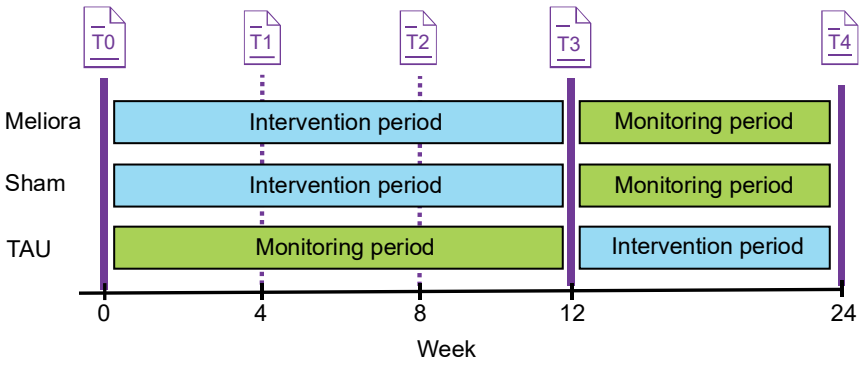
46. Dimidjian, S., Barrera, M., Martell, C., Muñoz, R. F. & Lewinsohn, P. M. The origins and current status of behavioral activation treatments for depression. Annu Rev Clin Psychol 7, 1–38 (2011).

47. Cuijpers, P., van Straten, A. & Warmerdam, L. Behavioral activation treatments of depression: A meta-analysis. Clin Psychol Rev 27, 318–326 (2007).

48. Alber, C. S., Krämer, L. V., Rosar, S. M. & Mueller-Weinitschke, C. Internet-Based Behavioral Activation for Depression: Systematic Review and Meta-Analysis. J Med Internet Res 25, e41643 (2023).

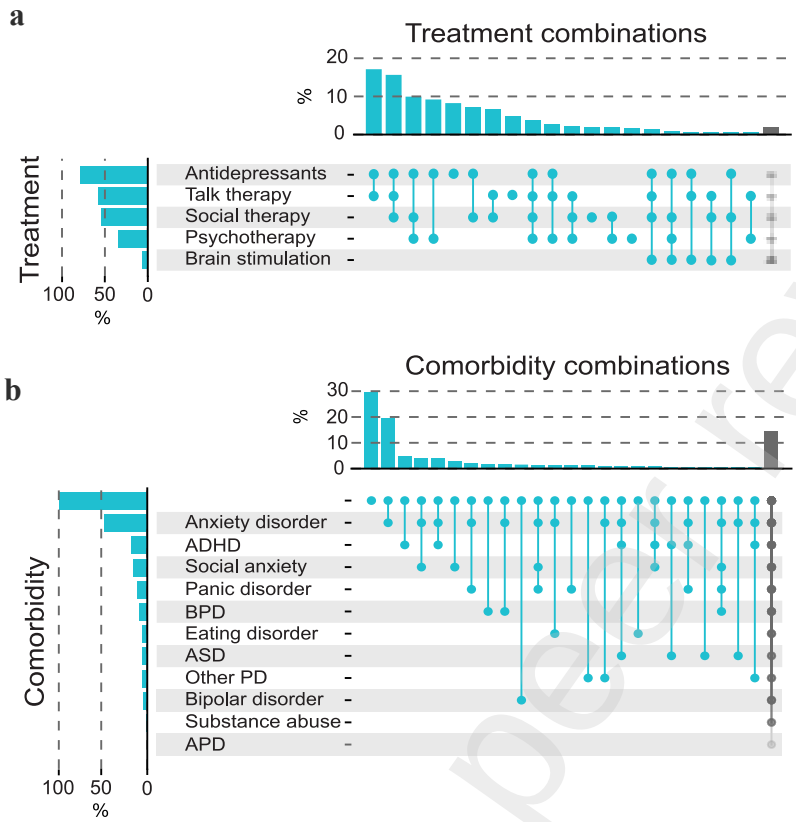
49. Cuijpers, P., Karyotaki, E., Harrer, M. & Stikkelbroek, Y. Individual behavioral activation in the treatment of depression: A meta analysis. Psychotherapy Research 33, 886–897 (2023).

Suppl. Fig. 1: Clinical trial design.



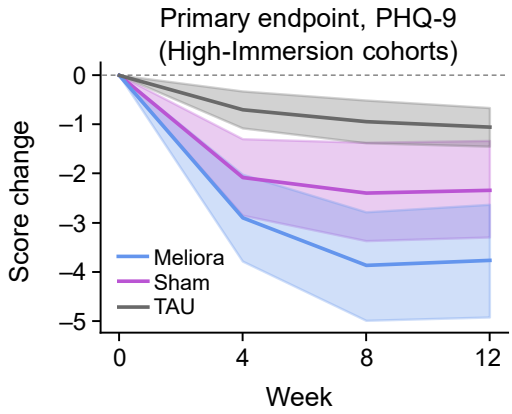
Patients in the MEL-T01 (active device, Meliora) and MEL-S01 (comparator device, Sham) arms underwent a 12-week intervention period, while the TAU arm received treatment-as-usual in a monitoring period. Following the initial 12 weeks, a crossover was implemented, extending follow-up to 24 weeks. Self-reported symptom scales were administered digitally at five time points (T0–T4) throughout the trial.

Suppl. Fig. 2: Baseline treatments and comorbidities.



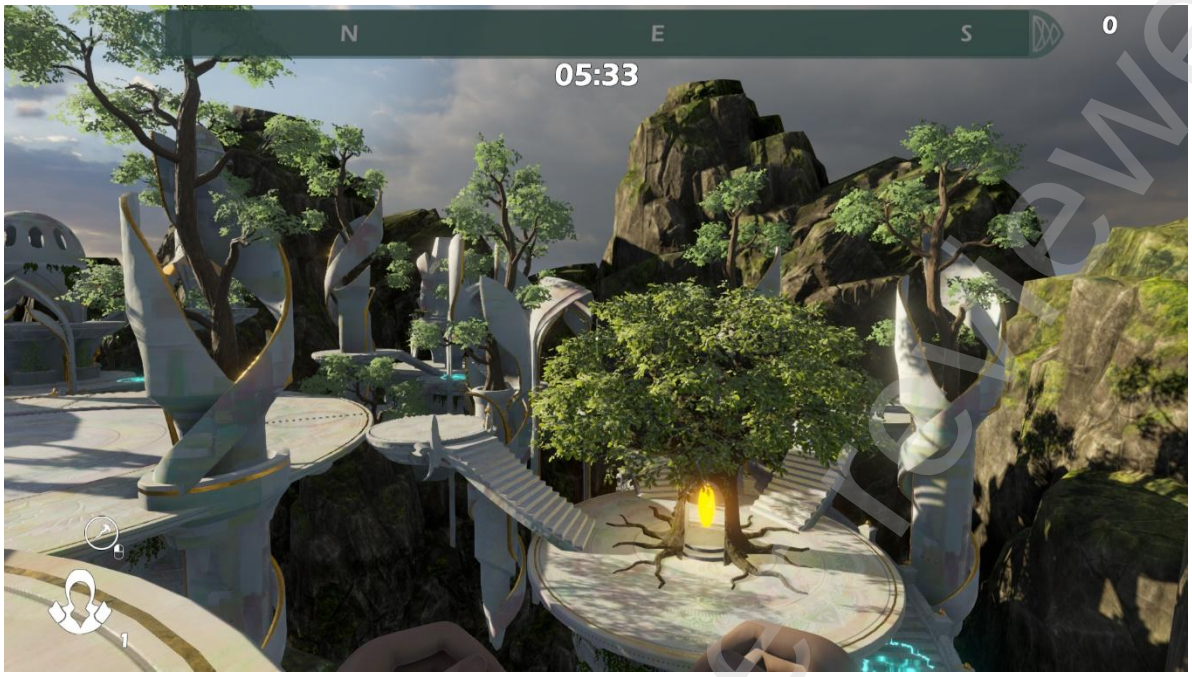
a, the baseline treatments and **b,** self-reported psychiatric comorbidities including the prevalence of individual conditions (left bar plots) and the presence of their combinations (right UpSet plots) in the per-protocol analysis population (n=483). BPD: Borderline personality disorder. ASD: Autism spectrum disorder. PD: Personality disorder. APD: Antisocial personality disorder.

Suppl. Fig. 3: Intervention efficacy in high-immersion cohorts.



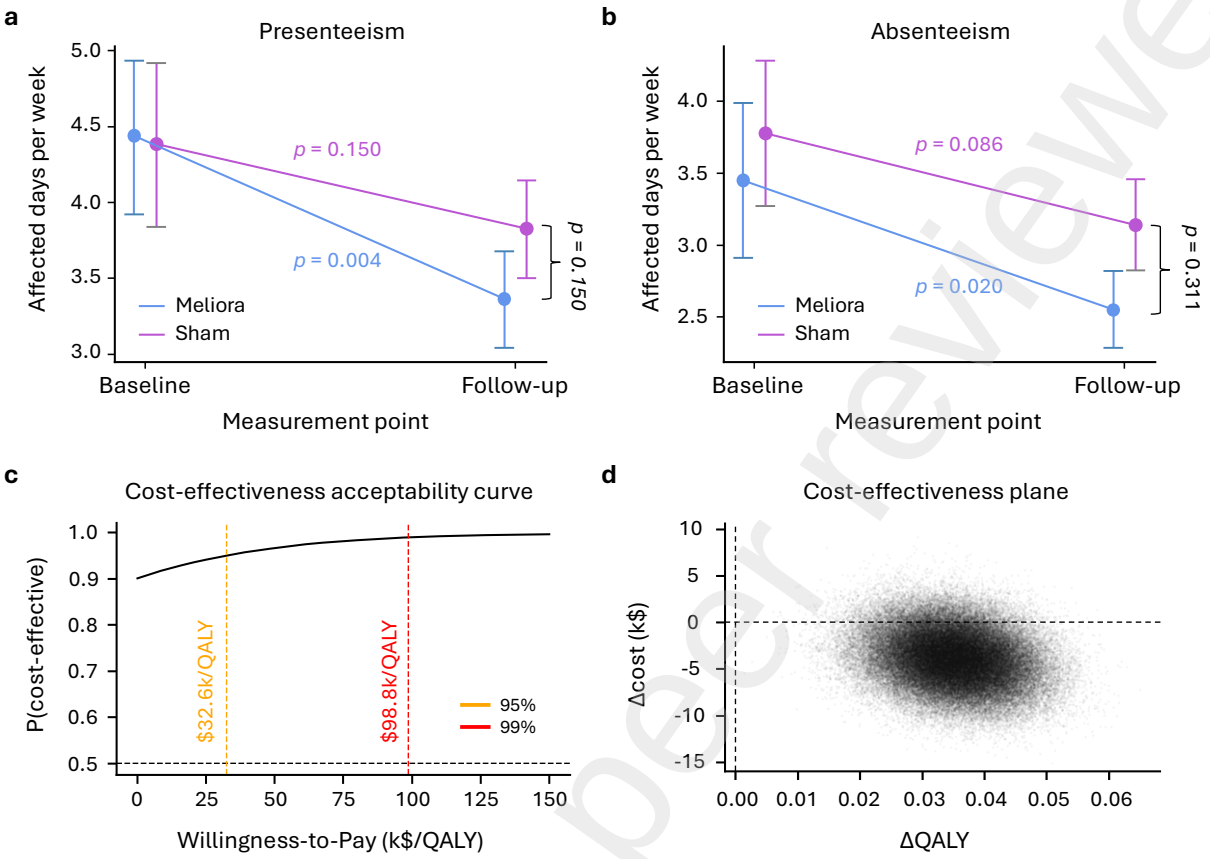
Time course of the PHQ-9 symptom change across treatment arms with above-median immersion scores. Shaded areas correspond to 5th and 95th bootstrapped confidence intervals of the mean. A split-cohort re-analysis of the primary outcome was performed by dividing the Meliora and Sham arms at the median of average self-reported immersion (IEQ). The estimated contrasts between groups are: Meliora vs. Sham: *c* –0·776, d –0.285, p=0.018; Sham vs. TAU: c –1·021, d –0·375, p=$2 \cdot 8 \cdot 10^{-4}$; Meliora vs. TAU: c –1·797, d –0·660, p=$1 \cdot 9 \cdot 10^{-10}$.

Suppl. Fig. 4: Intervention visual design.



Representative in-game visual design of the investigational device (MEL-T01, Meliora) and the active comparator (MEL-S01, Sham). The image illustrates the immersive virtual environment experienced by patients.

Suppl. Fig. 5: Cost-effectiveness analysis



a, Change in presenteeism (affected workdays/week) from baseline to follow-up in the Meliora and Sham groups. **b,** Change in absenteeism (affected workdays/week) from baseline to follow-up. Coloured p-values indicate within-group changes; vertical p-value reflects between-group differences. **c,** Cost-effectiveness acceptability curve showing the probability that Meliora is cost-effective across a range of willingness-to-pay thresholds. Vertical lines indicate 95% and 99% cost-effectiveness thresholds. **d,** Cost-effectiveness plane from bootstrap simulations, plotting incremental cost against incremental QALYs.

Suppl. Table 1: Digital intervention mechanisms of action in Meliora and Sham.

This table outlines key psychological and cognitive mechanisms embedded in the intervention design, categorized by concept and implementation feature. CBT = Cognitive Behavioral Therapy; EFT = Executive Functions Training. "Yes" indicates that the corresponding mechanism was implemented in the specified intervention arm.

Mechanism of action	Concept	Intervention feature	In-intervention implementation	Meliora	Sham
In-game activation	Treatment structure, style, and principles	Overall intervention design	Intervention's look and feel closely resembles commercial video games fostering positive experiences, enjoyment, and repeated use, while building a therapeutic relationship through context-sensitive tutorialization and support.	Yes	Yes
	Activity scheduling and structuring	Progress, goals, and feedback	The intervention is structured and time sensitive. The patient is given clear goals, feedback on their progress, and rewards for success, and they unlock narrative and powerful abilities through the 28 intervention levels for a sense of competence and achievement.	Yes	Yes
	Positive reinforcement	Within-level interactions	The patient controls and navigates a character in a complex 3-dimensional fictive environment with keyboard and mouse for sensorimotoric challenge.	Yes	Yes
CBT-based game narrative	Cognitive restructuring	Narrative	The patient progresses through a character-driven, voice-over, subtitled CBT-inspired narrative focusing on cognitive defusion and identifying cognitive biases.	Yes	Yes
EFT	Implicit cognitive training	Action-video-game experience	The patient overcomes challenging, high-paced tactical adversarial encounters while engaging in strategic area control task both offering high skill-ceiling.	Yes	Yes
	Explicit cognitive training	Targeted broad-spectrum training of executive functions	The intervention includes individually adaptive minigames and adversarial encounters enriched with further strategic decision-making and task-switching through ally non-player characters implementing broad-spectrum training of cognitive functions.	Yes	No

Suppl. Table 2: Baseline demographic variables by study arm.

Distributions of income, life status, education, handedness, and number of children are shown for the per-protocol cohort in each study arm (MEL-T01, MEL-S01, TAU). *p*-values indicate whether distributions differ significantly across groups.

Category	Variable	T01 (n=99)	S01 (n=96)	TAU (n=288)	p
Monthly income	< 1000€	41·4 %	34·4 %	33·7 %	
	1000–1999 €	30·3 %	31·2 %	26·7 %	
	2000–2999 €	14·1 %	15·6 %	14·2 %	
	3000–3999 €	7·1 %	8·3 %	12·2 %	
	4000–4999 €	2·0 %	2·1 %	5·6 %	0·66
	> 5000 €	1·0 %	3·1 %	1·7 %	
	Prefer not to say	4·0 %	5·2 %	5·9 %	
Life status	Unemployed	9·1 %	10·4 %	12·5 %	
	Retired	3·0 %	2·1 %	0 %	
	Partial or disability pension	3·0 %	6·3 %	5·9 %	
	Sick leave or rehabilitation support	29·3 %	26·0 %	20·8 %	
	Employed part-time	13·1 %	7·3 %	10·1 %	0.30
	Student	23·2 %	26·0 %	25·0 %	
	Parental leave	0 %	0 %	0·4 %	
	Employed full-time	19·2 %	21·9 %	25·4 %	
Education	Primary school	8·1 %	4·2 %	10·1 %	
	High school	40·4 %	31·2 %	28·1 %	
	Vocational education	21·2 %	26·0 %	24·0 %	
	Bachelor's degree	18·2 %	25·0 %	21·2 %	
	Master's degree	11·1 %	11·5 %	15·3 %	0·55
	Licenciate	0 %	0 %	0·3 %	
	Doctoral degree	1·0 %	2·1 %	1·0 %	
Handedness	Right-handed	87·9 %	88·5 %	92·0 %	
	Left-handed	5·1 %	7·3 %	5·6 %	0.29
	Ambidextrous	7·1 %	4·2 %	2·4 %	
Children	0	77·8 %	74·0 %	76·0 %	
	1	8·1 %	10·4 %	9·7 %	
	2	8·1 %	8·3 %	9·4 %	0·96
	3	4·0 %	5·2 %	2·8 %	
	4 or more	2·0 %	2·1 %	2·0 %	

Suppl. Table 3: Baseline treatment contact.

Distribution of self-reported healthcare contacts across the three study arms (MEL-T01, MEL-S01, TAU) in the per-protocol cohort. *p*-values indicate whether the distribution of contact types differs between groups.

Category	Variable	T01 (n=99)	S01 (n=96)	TAU (n=288)	*p*
Treatment contact	Specialized healthcare	41·4 %	44·8 %	30·6 %	0·28
	Primary healthcare	10·1 %	6·2 %	13·5 %	
	Student healthcare	14·1 %	18·8 %	15·3 %	
	Occupational healthcare	27·3 %	26·0 %	27·8 %	
	Private healthcare	14·1 %	14·6 %	18·1 %	
	Third sector	4·0 %	2·1 %	1·4 %	

Suppl. Table 4. Cognitive training effects and their relationship with symptom score change.

Within-patient improvements across planning, memory, processing speed, and cognitive control, reported as standardized slopes (*s*), percent change, and standardized effect sizes (*d'*). Correlations (*r*) between cognitive gains and symptom change are shown across clinical outcomes (PHQ-9, QIDS-SR16, GAD-7, RRS-SV, SDS, WHO-5, PVSS-SF), with associated *p*-values. Significant associations are highlighted. Improvements were observed in planning time, working memory and short-term memory retrieval, processing speed, long-term memory, and cognitive control. Several of these improvements were significantly associated with reductions in depression, anxiety, and functional impairment. Specifically, planning time was associated with symptom reductions across PHQ-9, QIDS-SR16, and SDS; working memory retrieval with PHQ-9, GAD-7, and SDS; processing speed with PHQ-9 and SDS; short-term memory retrieval with SDS; and long-term memory retrieval with GAD-7. See Methods for task descriptions and analysis details.

						Correlation of cognitive improvement with reduction of symptoms					
Construct	**Measure**	**Slope**	**Change (%)**	**Change *d'***	**PHQ-9**	**QIDS-SR16**	**GAD-7**	**RRS-SV**	**SDS**	**WHO-5**	**PVSS-SF**
Planning	Planning time	s = -0.004 $p = 4.9 \cdot 10^{-8}$**	-22·3	–0·85	r = 0·234 p = 0·011*	r = 0·203 p = 0·024*	r = 0·152 p = 0·070	r = 0·016 p = 0·440	r = 0·213 p = 0·018*	r = –0·136 p = 0·907	r = –0·012 p = 0·545
	Planning efficiency	$s = 1 \cdot 7 \cdot 10^{-4}$ $p = 4 \cdot 3 \cdot 10^{-4}$	3·37	0·47	r = 0·040 p = 0·702	r = 0·018 p = 0·569	r = 0·092 p = 0·814	r = 0·195 p = 0·972	r = –0·009 p = 0·534	r = –0·002 p = 0·492	r = –0·123 p = 0·117
Working memory	Encoding time	$s = 1 \cdot 9 \cdot 10^{-4}$ p = 0·772	1·92	0·04	r = –0·004 p = 0·484	r = 0·049 p = 0·317	r = 0·037 p = 0·360	r = –0·056 p = 0·292	r = 0·111 p = 0·138	r = –0·037 p = 0·641	r = –0·001 p = 0·502
	Retrieval time	s = -0·002 $p = 1 \cdot 7 \cdot 10^{-14}$**	-15·86	–1·25	r = 0·207 p = 0·020*	r = 0·153 p = 0·065	r = 0·250 p = 0·006*	r = –0·054 p = 0·296	r = 0·239 p = 0·008*	r = –0·209 p = 0·981	r = –0·099 p = 0·836
	Retrieval error	$s = -2 \cdot 3 \cdot 10^{-5}$ p = 0·226	7·75	0·17	r = –0·027 p = 0·396	r = –0·001 p = 0·496	r = 0·131 p = 0·098	r = 0·010 p = 0·460	r = –0·043 p = 0·338	r = 0·041 p = 0·655	r = 0·037 p = 0·641
Processing speed	Execution time	$s = -9 \cdot 6 \cdot 10^{-4}$ $p = 1 \cdot 1 \cdot 10^{-5}$**	–25·7	–0·67	r = 0·243 p = 0·008*	r = 0·109 p = 0·144	r = 0·155 p = 0·066	r = 0·041 p = 0·346	r = 0·198 p = 0·026*	r = 0·030 p = 0·613	r = 0·014 p = 0·555
Short-term memory	Retrieval time	s = –0·025 $p = 1 \cdot 4 \cdot 10^{-5}$**	–33·9	–1·21	r = 0·163 p = 0·058	r = 0·140 p = 0·088	r = 0·087 p = 0·203	r = –0·073 p = 0·241	r = 0·267 p = 0·004*	r = –0·128 p = 0·891	r = –0·189 p = 0·966
	Retrieval error	s = –0·002 $p = 4 \cdot 4 \cdot 10^{-4}$**	–46·2	–0·51	r = 0·008 p = 0·468	r = 0·017 p = 0·436	r = –0·130 p = 0·108	r = –0·145 p = 0·082	r = 0·067 p = 0·262	r = 0·019 p = 0·570	r = 0·057 p = 0·705
Long-term memory	Retrieval time	s = –0·079 p = 0·007**	–50·5	–0·40	r = 0·130 p = 0·108	r = –0·028 p = 0·396	r = 0·214 p = 0·002*	r = 0·071 p = 0·248	r = 0·084 p = 0·213	r = –0·140 p = 0·910	r = –0·104 p = 0·840
	Retrieval error	s = 0·003 p = 0·436	–12·9	–0·11	r = 0·080 p = 0·222	r = 0·107 p = 0·153	r = 0·050 p = 0·318	r = –0·089 p = 0·196	r = 0·111 p = 0·144	r = –0·122 p = 0·880	r = –0·068 p = 0·743
Cognitive control	Set-switch error	s = -0·003 $p = 2 \cdot 6 \cdot 10^{-8}$**	–49·7	–0·80	r = 0·037 p = 0·360	r = 0·074 p = 0·238	r = 0·052 p = 0·309	r = –0·120 p = 0·124	r = 0·122 p = 0·120	r = –0·108 p = 0·849	r = –0·086 p = 0·796

Suppl. Table 5: Adverse Events across Meliora and Sham devices.

Adverse events (AEs) are reported separately for the investigational device (MEL-T01), active comparator (MEL-S01), and pooled across both devices. Counts and percentages reflect the number of patients reporting each event category. AE categories are grouped by type, severity, relatedness, expectedness, and whether the event led to self-reported discontinuation. No severe AEs were reported in either group.

Category	Sub-category	T01 (*n*=490)		S01 (*n*=511)		T01+S01 (*n*=1,001)	
		Count	%	Count	%	Count	%
One AE or more		**77**	**15·7**	**71**	**13·9**	**148**	**14·8**
AE categories	Frustration	43	8·8	44	8·6	87	8·7
	Stress	17	3·5	16	3·1	33	3·3
	Anxiety	13	2·7	15	2·9	28	2·8
	Nausea	10	2·0	11	2·2	21	2·1
	Headache	4	0·8	2	0·4	6	0·6
	Arm and neck pain	3	0·6	2	0·4	5	0·5
	PTSD exacerbates	2	0·4	1	0·2	3	0·3
	Audiovisual discomfort	0	0·0	3	0·6	3	0·3
	Nightmares	1	0·2	0	0·0	1	0·1
Severity	Mild	75	15·3	71	13·9	146	14·6
	Moderate	4	0·8	4	0·8	8	0·8
	Severe	0	0·0	0	0·0	0	0·0
Relatedness	Definitive	73	14·9	67	13·1	140	10·0
	Probable	6	1·2	5	1·0	11	1·1
	Possible	3	0·6	4	0·8	7	0·7
Expectedness	Expected	65	13·3	64	12·5	129	12·9
	Unexpected	15	3·1	17	3·3	32	3·2
Discontinuation	No self-reported discontinuation	69	14·1	65	12·7	134	13·4
	Self-reported discontinuation	11	2·2	8	1·6	19	1·9

Suppl. Table 6: Model residual diagnostics for linear mixed models (LMMs).

Shapiro–Wilk tests were used to assess the normality of residuals, and Levene's tests were used to assess homogeneity of variance across groups. Significant violations were observed for several endpoints, supporting the use of robust LMMs in the main analyses. * indicates p<0·05

Symptom Scale	Shapiro-Wilk *p*-value	Levene's test *p*-value
PHQ-9	$5{\cdot}4{\cdot}10^{-7}$ *	0·012 *
QIDS-SR16	$2{\cdot}6{\cdot}10^{-8}$ *	0·447
GAD-7	0·006 *	0·426
SDS	$2{\cdot}6{\cdot}10^{-8}$ *	0·190
RRS-SV	$1{\cdot}8{\cdot}10^{-13}$ *	0·623
WHO-5	$8{\cdot}5{\cdot}10^{-17}$ *	0·462
PVSS-SF	$1{\cdot}4{\cdot}10^{-12}$ *	0·654