Multi-site randomised controlled trial of outcome feedback technology used to support the psychological treatment of depression and anxiety

Jaime Delgadillo[1], Kim de Jong[2], Mike Lucock[3,12], Wolfgang Lutz[4], Julian Rubel[4], Simon Gilbody[5], Shehzad Ali[5], Elisa Aguirre[6], Mark Appleton[7], Jacqueline Nevin[8], Harry O'Hayon[9], Ushma Patel[10], Andrew Sainty[11], Peter Spencer[12], and Dean McMillan[5]

1. Clinical Psychology Unit, Department of Psychology, University of Sheffield, Sheffield, United Kingdom
2. Institute of Psychology, Leiden University, The Netherlands
3. Centre for Applied Research in Health, University of Huddersfield, Huddersfield, United Kingdom
4. Department of Psychology, University of Trier, Germany
5. Department of Health Sciences, University of York, United Kingdom
6. North East London NHS Foundation Trust, London, United Kingdom
7. Pennine Care NHS Foundation Trust, Hyde, United Kingdom
8. Cheshire and Wirral Partnership NHS Foundation Trust, Cheshire, United Kingdom
9. Whittington Health NHS Trust, London, United Kingdom
10. Cambridgeshire and Peterborough NHS Foundation Trust, Cambridge, United Kingdom
11. Humber NHS Foundation Trust, Hessle, United Kingdom
12. South West Yorkshire Partnership NHS Foundation Trust, Barnsley, United Kingdom

Declaration of interests: None.

[1] Correspondence: Dr Jaime Delgadillo, Clinical Psychology Unit, University of Sheffield, Floor F, Cathedral Court, 1 Vicar Lane, Sheffield S1 2LT, UK. jaime.delgadillo@nhs.net

Summary

Background: Previous research suggests that using outcome feedback technology can enable psychological therapists to identify and resolve obstacles to clinical improvement. This study aimed to evaluate the effectiveness of an outcome feedback quality assurance system applied in stepped care psychological services.

Methods: This multi-site cluster randomised controlled trial (registration DOI: 10.1186/ISRCTN12459454) included 2233 patients with depression and anxiety disorders accessing at least 2 sessions of individual psychological therapy delivered by 77 therapists across 8 healthcare organisations. Therapists were randomised to a feedback intervention group (N = 39) or a treatment-as-usual control group (N = 38). The feedback technology alerted therapists to cases that were "*not on track*", and primed them to review these in clinical supervision. Post-treatment symptom severity on validated depression (PHQ-9) and anxiety (GAD-7) measures was compared between groups using multilevel modelling, controlling for cluster (therapist) effects, following an *intention-to-treat* approach.

Findings: Cases classified as *not on track* had significantly less severe symptoms after treatment if they were allocated to the feedback group (PHQ-9 d = 0.23, B = -1.03 [95% CI = -1.84, -0.23], p = 0.012; GAD-7 d = 0.19, B = -0.85 [95% CI = -1.56, -0.14], p = 0.019). There were no between-group differences in the odds of reliable improvement (OR = 1.32 [0.93, 1.89], p = 0.12); however, control cases classed as *not on track* had significantly greater odds of reliable deterioration (OR = 1.73 [1.18, 2.54], p = 0.0050).

Interpretation: Supplementing psychological therapy with low-cost feedback technology prevents deterioration in cases at risk of poor response to treatment. This evidence supports the implementation of outcome feedback in stepped care psychological services.

Research in context

Evidence before this study

Previous research suggests that using inexpensive quality improvement strategies such as routine outcome monitoring and feedback can improve psychological treatment outcomes, in particular for cases that are prone to deterioration. The generalisability of previous trials is limited by their application in specialist university or psychotherapy clinics, and observational studies in primary care were likely to be statistically underpowered.

Added value of this study

This large-scale, pragmatic, randomised controlled trial was adequately powered to detect small effect size differences, and designed to evaluate the generalisability of feedback effects across multiple primary care psychological therapy services.

Implications of all the available evidence

There is now a compelling evidence base to support the implementation of outcome monitoring and feedback technologies in mainstream psychological services. Implementing this low-cost, automated feedback and quality assurance system can help to prevent deterioration for cases that are at risk of poor treatment outcomes.

Introduction

A number of psychological interventions, ranging from brief guided self-help to more intensive psychotherapies, are effective for the treatment of depression and anxiety disorders.[1] Large-scale evaluations of such treatments applied in routine care are generally favourable, although it is also known that at least 30% of patients do not show statistically reliable improvement and some deteriorate.[2-3] Previous studies have shown that patients at risk of poor response to treatment can be identified early using outcome feedback methods.[4] Outcome feedback is a quality assurance method which involves routinely monitoring a patient's condition using standardised measures which are compared to data from a normative clinical sample.[5] Using data charts or automated electronic monitoring technologies, cases that are "*not on track*" are detected when their symptoms are significantly worse than those of similar cases.

Several reviews of experimental and practice-based studies suggest that using outcome feedback can help to improve treatment outcomes by comparison to usual psychological care.[4,6-8] Simply collecting patient-reported outcome measures in clinical practice is not associated with improved outcomes,[9] so it is plausible that the "risk signal" element of feedback technologies serves to effectively prompt therapists to identify and to resolve obstacles to improvement. This mechanism of action is supported by evidence from controlled trials where therapy supported with risk signalling yielded better outcomes than routine psychological care.[6-8] An early meta-analysis suggested that supplementing the signal with clinical decision-making and support tools further enhances its effectiveness,[6] although a more recent meta-analysis contradicts this finding.[9] It has also been proposed that outcome feedback specifically helps to prevent deterioration in cases classed as *not on track*.[6-8] A recent systematic review of the literature concluded that studies that applied risk signalling technology show some

evidence of improved outcomes for *not on track* cases, but the effect sizes were small (standardised mean difference of -0.22).[9] Furthermore, some studies have not found a differential effect of feedback in the *not on track* subgroup[10-12] and one study found that using feedback possibly deteriorates outcomes for *not on track* cases with cluster B personality disorders.[13]

Overall, the literature shows mixed and inconclusive evidence for the use of feedback technologies, and the methodological quality of studies has been rated as generally low.[9] This variability raises questions about the generalisability of feedback, justifying the need to carefully evaluate its acceptability, feasibility and effectiveness prior to adoption in routine care.[14] Some studies have suggested that outcome feedback may be particularly helpful in short-term evidence-based therapies such as cognitive behavioural therapy, and could enhance the efficiency of treatment.[10,11] A recent study reported qualitative evidence that feedback-assisted brief psychological interventions were acceptable to patients with depression and anxiety disorders, and feasible to implement in a routine primary care setting.[11] This study also suggested that outcome feedback could reduce the cost and enhance the efficiency of treatment, although it was limited by the use of historical control group data in a non-randomized design. In spite of these promising results, more rigorous experimental evidence is necessary to establish the generalisability and efficacy of feedback in primary care settings. The present study aimed to address this gap in the literature through a multi-site randomised controlled trial applied in primary care psychological services for common mental health problems.

Methods

Study design

This was a pragmatic, multi-site, cluster randomised controlled trial. The objective was to assess the clinical effect of feedback-assisted psychological treatments, in comparison to routinely delivered psychological care. The central hypothesis was that using feedback would result in lower mean symptom severity for *not on track* cases, in comparison to usual care. The primary outcome was depression and anxiety symptom severity assessed at the last treatment session using validated patient-reported outcome measures described below. Secondary outcomes included work and social adjustment, treatment duration, reliable improvement, reliable deterioration, treatment dropout rates and the percentage of cases classified as *not on track.*

The design involved randomising participating therapists (and all of their patients meeting inclusion criteria described below) to a feedback intervention group or a treatment-as-usual control group. The rationale for this design was two-fold. First, randomising therapists would minimise the risk of contamination of controls through practice effects, which could occur if the same therapist were to treat some patients with and others without using outcome feedback technology. Secondly, this cluster design adequately represents the natural nesting of patients within therapists, thus enabling us to control for variability in outcomes attributable to therapists (*therapist effects*[15]).

Using the Optimal Design Software for Multi-level and Longitudinal Research (Version 3.01)[16], we estimated that a minimum of 60 therapists (30 per group) –each of whom treated an average of 10 patients– was required to detect a small effect size with an alpha level of $\alpha = 0.05$ and 80% power. This calculation assumed an intracluster correlation coefficient of ICC = 0.05, guided by previous

studies investigating *therapist effects* in naturalistic samples.[15,17] We aimed to recruit up to 80 therapists to account for attrition.

The study was approved by the London - City & East NHS Research Ethics Committee (06/01/2016, Ref: 15/LO/2200) and the protocol was registered in an international database prior to recruitment (DOI: 10.1186/ISRCTN12459454).

Setting and interventions

The study was conducted in eight National Health Service (NHS) Trusts in England. Together, these services covered a large primary care population across London, Cambridge, Cheshire & Wirral, Bury, Heywood, Middleton, Rochdale, Oldham, Stockport, Tameside & Glossop, Trafford, Barnsley, and East Riding.

All participating services were part of the national *Improving Access to Psychological Therapies* (IAPT) programme, which offers protocol-driven, evidence-based psychological interventions for depression and anxiety disorders organised in a stepped care model.[18] Low intensity guided self-help based on principles of cognitive behavioural therapy (LiCBT) was offered as an initial treatment in most cases with mild-to-moderate depression and/or anxiety problems. LiCBT is delivered by trained coaches (*psychological wellbeing practitioners*) in a variety of different formats (e.g., individual or group psychoeducation, computerised CBT with telephone support) and typically lasts under 8 sessions. Those with more severe or complex problems, and those who did not respond to LiCBT were "stepped up" to high intensity (up to 20 sessions) psychotherapies including CBT, interpersonal psychotherapy, and counselling for depression. The specific treatment recommendation for each case followed standard clinical guidelines.[19] Treatment was supported by regular (weekly or fortnightly) clinical supervision delivered in a peer-supervision model organised within each service.

Participants

Therapists qualified to deliver low or high intensity interventions were eligible to take part, with the exception of (1) therapists with short-term employment contracts or (2) trainees who were not yet fully qualified. The trial included all patients that accessed individual (low and/or high intensity) therapy with participating therapists, excluding patients who accessed group therapies and those who attended less than 2 individual therapy sessions. The latter condition was applied because: (1) outcome measures for patients that accessed 1 session reflect symptom severity for a pre-treatment period of 2 weeks, and (2) the outcome feedback technology requires at least 2 sessions to provide a progress feedback signal taking session 1 as a baseline score. The allocation of patients to therapists in routine care was quasi-random, where patients on waiting list were allocated sequentially based on therapist availability.

Outcome feedback quality assurance system

Therapists in all participating services routinely recorded their patients' clinical outcomes using an electronic clinical record system called *Patient Case Management Information System* (PCMIS; http://www.pc-mis.co.uk). PCMIS includes outcome monitoring graphs which chart depression and anxiety symptom severity scores at every session. Therapists randomised to the experimental group had access to enhanced outcome monitoring graphs which included *expected treatment response* curves. The *expected treatment response* curves represent 80% prediction intervals, which are estimated using growth curve modelling in data from a normative clinical sample.[5,20-21] *Expected treatment response* curves were calculated for subgroups of cases with the same baseline symptom severity, using a large clinical dataset of cases treated in IAPT (further details described elsewhere[22]). These enhanced outcome monitoring graphs automatically generated

a “red signal” to alert therapists to *not on track* cases whose depression and/or anxiety symptoms surpassed the 80% upper boundary of the *expected treatment response* curves. Control group therapists only had access to standard outcome tracking graphs, but without *expected treatment response* curves or automated risk signals.

Therapists randomised to the feedback group attended a standardised 6.5-hour training programme which covered: outcome feedback theory and evidence-base; instructions on how to use the feedback tool; clinical trouble-shooting skills. The training required therapists to follow the following process: (1) review outcome feedback graphs with patients at the start of every session; (2) if the graph shows a risk signal, discuss this with the patient to collaboratively identify potential obstacles to improvement; (3) prioritise discussing *not on track* cases with your clinical supervisor; (4) use information from points 2 and 3 to develop a plan to address obstacles; (5) use outcome feedback graphs to assess how your plan is working. Therapists were also primed to be aware of variables that have been empirically shown to be associated with treatment outcomes (patient, therapist, process, and wider context factors). This information and evidence-base was synthesised in a clinical guideline that therapists assigned to the feedback group received after training.[23]

Outcome measures and secondary data

Patients accessing the participating services routinely self-completed standardised outcome measures before each session; the measures obtained at the last treatment session were taken as primary outcomes in the trial. The Patient Health Questionnaire (PHQ-9) is a nine-item screening tool for depression, where each item is rated on a 0 to 3 scale, yielding a total depression severity score between

0–27.[24] A cut-off ≥ 10 has been recommended to screen for major depression,[24] and a difference of ≥6 points between assessments is indicative of reliable change.[25]

The Generalized Anxiety Disorder questionnaire (GAD-7) is a seven-item measure developed to screen for anxiety disorders.[26] It is also rated using a 0 to 3 scale, yielding a total anxiety severity score between 0–21. A cut-off score ≥8 is recommended to identify the likely presence of a diagnosable anxiety disorder,[26] and a difference of ≥5 points is indicative of reliable change.[25]

Secondary data sources included demographics (age, gender, ethnicity, employment status), stepped care pathway information, number of treatment sessions, primary diagnoses recorded in clinical records and functional impairment measured using the Work and Social Adjustment Scale (WSAS).[27]

Recruitment, randomisation and data collection

Recruitment took place between January and July 2016. A participant information sheet and consent form were shared via email with all therapists working in participating services. Therapists had an opportunity to clarify questions with the principal investigator before providing signed consent forms directly to the research team. Parallel-group random allocation was independently performed by a researcher using a computer-generated (1:1) randomisation algorithm to prevent selection bias within services. Given the nature of the outcome feedback technology, this was an open-label trial where therapists were aware of their allocation. Session-by-session depression (PHQ-9) and anxiety (GAD-7) outcome measures were collected for all eligible patients who accessed individual therapy with participating therapists during a one-year study period.

Data analysis

Patients' characteristics were compared between groups (those included and excluded from the trial sample) using Mann-Whitney U tests for continuous variables and chi-square tests for categorical variables. A small number of cases (N = 98; 4.4% of the trial sample) had missing post-treatment outcome measures which were imputed by averaging the imputed values from 25 estimated datasets using an expectation maximization method.[28] This imputation was carried out so that we could conduct *intention-to-treat* analyses, including post-treatment outcomes for all cases regardless of completion or dropout status.

The primary analysis was carried out using multilevel modelling (MLM) with separate models for PHQ-9 and GAD-7 outcomes. Following conventional model building guidelines,[29] we initially examined the hierarchical structure of the dataset using unconditional models predicting post-treatment symptom severity. The "site" variable was not statistically significant in a three-level model (patients within therapists within sites), so subsequent analyses used two-level models (patients within therapists). Next, we considered different covariance structures, assessed non-linear (i.e., quadratic, log-linear) trends in the number of treatment sessions, and assessed goodness-of-fit (using AIC, BIC, −2 log likelihood statistics). After initial model checking, the primary analysis applied a two-level model, including random intercepts for therapists, with an unstructured covariance matrix, and an identity link-function. No cases included in the trial sample had two interventions delivered by different therapists (e.g., low followed by high intensity therapy), so crossed random effects were not modelled. Continuous variables were grand-mean centred and an intracluster correlation coefficient (ICC) was calculated to assess the proportion of variance in outcomes attributable to therapists. An initial conditional model included the following predictors: *baseline severity* of symptoms, log transformed number of *sessions*, and *group* (feedback

vs. control), which compared between-group differences in post-treatment symptom severity. Next, a fully adjusted model additionally included a case *classification* (case classified as *on track* vs. *not on track*), and a *group* * *classification* interaction term (main hypothesis test). This MLM strategy was repeated in a sensitivity analysis controlling for age and step of care (low vs. high intensity treatment).

Secondary analyses assessed other relevant clinical outcomes. The fully adjusted MLM was repeated using the WSAS as a dependent variable to assess potential effects of feedback on functional impairment. Poisson MLM was used to compare between-group differences in treatment duration, controlling for baseline PHQ-9 and GAD-7. Logistic MLM was used to compare between-group probabilities (odds ratios) of meeting post-treatment criteria for reliable improvement (RI), after controlling for baseline severity (PHQ-9 and GAD-7). The RI classification required patients to have statistically reliable improvement in at least one of the outcome measures, as long as the other measure did not show reliable deterioration. Logistic MLM was also used to estimate between-group odds ratios for the % of cases with reliable deterioration (in at least one outcome measure), the percentage of cases classed as *not on track*, and the percentage of cases that dropped-out of treatment. These models were computed using the full sample and repeated in the *not on track* subsample.

Role of the funding source

The study was partly supported by research capability funding awarded by the English National Health Service (NHS) and partly funded by a visiting research fellowship awarded to the principal investigator by the Department of Health Sciences, University of York. The funding organisations had no role in the decision to publish the study.

Results

Sample characteristics

In total, 79 therapists were recruited but 2 did not participate (see Figure 1). Of the 77 participating therapists, 39 (50.6%) were randomised to the feedback group and 38 (49.4%) to the control group. Of these, 48 (62.3%) delivered high intensity CBT, 23 (29.9%) delivered low intensity CBT, and 6 (7.8%) delivered counselling for depression. Most therapists were females (84.4%) from a white British background (84.4%), with an average of 7 years' experience in delivering psychological interventions (range = 9 months to 31 years). The number of trial cases treated by each therapist ranged between 1 and 113 (median = 25, mean = 30.77, SD = 24.54). Further sample characteristics are summarised in Table 1.

Altogether, 2233 patients meeting case selection criteria described above were included in the trial (1176 feedback cases, 1057 controls). According to clinical records, 34.5% had a primary affective disorder (major depression episode, recurrent depression), 14.2% had mixed anxiety and depression disorder, 14.6% had generalized anxiety disorder, 6.0% had post-traumatic stress disorder, and other anxiety problems were less prevalent. The mean number of weekly therapy sessions was 6.45 (SD = 3.67, median = 6, range = 2 to 25) in the full study sample; 6.35 (SD = 3.60, median = 6, range = 2 to 25) in the control group and 6.54 (SD = 3.73, median = 6, range = 2 to 22) in the OF group. Demographics and clinical characteristics are summarised in Table 1.

The trial sample excluded 651 cases that did not access individual therapy (e.g., group psycho-education cases) or who only attended a single session. Excluded cases had similar baseline characteristics compared to trial cases, but a higher proportion of unemployed patients (22.3% vs. 18.1%; p = 0.040) and marginally higher baseline PHQ-9 scores (mean difference = 0.35; p = 0.007).

[Figure 1]

[Table 1]

Primary analysis

The main effect for trial *group* was not statistically significant in the initial conditional models testing between-group differences (shown in supplementary appendix), nor in the fully adjusted models testing interaction terms (shown in Table 2). The negative coefficients for the *group * classification* interaction terms indicated that *not on track* cases tended to have lower post-treatment symptoms if they were in the feedback group, as depicted in Figure 2. The interaction was statistically significant in the depression model (B = -1.03, SE = 0.41, p = 0.012), and in the anxiety model (B = -0.85, SE = 0.36, p = 0.019). Approximately 11% of variability in depression (ICC = 0.107) and anxiety (ICC = 0.114) outcomes was attributable to *therapist effects*. Effect size differences between groups were PHQ-9 d = 0.17 and GAD-7 d = 0.13 in the whole sample (N = 2233); the corresponding values in the *not on track* subsample (N = 1288) were PHQ-9 d = 0.23 and GAD-7 d = 0.19. Sensitivity MLM analyses controlling for age and intensity of treatment (low vs. high) confirmed the same results (see supplementary appendix).

[Figure 2]

[Table 2]

Secondary analyses

The fully adjusted MLM results using WSAS as a dependent variable followed the same pattern as described above. The main effect for *group* was not significant (B = 0.46, SE = 0.77, p = 0.55), but the *group* * *classification* interaction term was statistically significant (B = -1.75, SE = 0.62, p = 0.0050) yielding an effect size of d = 0.22 in the *not on track* subgroup. The poisson MLM results indicated no significant differences in treatment duration between groups (B = -0.05, SE = 0.05, p = 0.37); and no significant *group * classification* interaction (B = -0.02, SE = 0.04, p = 0.62). Full outputs from these MLM analyses are in the supplementary appendix.

Table 3 summarises indices of clinical effectiveness. MLM results controlling for *therapist effects* indicated that there were no significant between-group differences in the odds of reliable improvement in the full sample (OR = 1.21, p = 0.29) or in the *not on track* subsample (OR = 1.32, p = 0.12). However, control cases had greater odds of reliable deterioration (full sample OR = 1.48, p = 0.023; *not on track* subsample OR = 1.73, p = 0.0050). There were no significant between-group differences in the odds of treatment dropout or of being classed as *not on track*.

[Table 3]

Discussion

Findings in context

This large-scale, multi-site trial conducted in stepped care IAPT services demonstrated that using low-cost outcome feedback technology can improve outcomes for cases that are at risk of poor response to treatment. No main effect of feedback was found overall; instead an interaction effect indicated that feedback is specifically helpful for cases classified as *not on track*. These findings are largely

consistent with reviews and meta-analyses of previous trials in university and outpatient psychotherapy centres, which conclude that the effect of feedback is mostly observed in *not on track* cases,[4,6,8,9] although there are also exceptions such as the trial by Amble et al.[12] which found main effects for feedback in the full sample but not in the *not on track* subgroup. Effect sizes of d = 0.23 for depression, d = 0.19 for anxiety, and d = 0.22 for work and social adjustment favouring the feedback group were observed. These effect sizes are small by conventional standards, but nevertheless remarkable considering the automated nature of the risk signalling technology and the low cost incurred by services in requiring outcome feedback users to attend a single-day training session. In addition, given that the feedback intervention prioritises clinical supervision resources for *not on track* cases, it is important to highlight that this did not disadvantage the *on track* cases in terms of clinical outcomes or dropout rates. Overall, this low-cost quality assurance system effectively integrates the use of routine outcome measures, outcome prediction technology and clinical supervision.

Given that usual treatment in IAPT stepped care services utilises standard outcome tracking charts and regular clinical supervision, we might expect modest effect size differences when supplementing this with risk signalling technology. Usual care (control) cases had higher rates of deterioration compared to feedback cases, although the odds ratios in this trial (full sample OR = 1.48; *not on track* subsample OR = 1.73) were lower by comparison to the OR = 2.3 reported in the meta-analysis by Shimokawa et al.[6] This difference may be influenced by the low base rate of cases with reliable deterioration in the participating services (<7.5%), whereas other psychotherapy settings have typically observed deterioration rates in the order of 10%.[1] This is plausibly explained by differences in case-mix, since IAPT services mostly support people with mild-to-moderate mental health problems.[18]

Contrary to recent studies applying evidence-based CBT interventions,[10-11] we found no significant effects of feedback on treatment duration. One methodological explanation could be that prior quasi-experimental studies did not have contemporaneous controls, and their effects on duration could be explained by other unmeasured factors that changed over time. An alternative explanation could be that the inclusion of counselling and LiCBT interventions in the present trial may have obscured effects that may be specific to conventional CBT. The potential influence of feedback on treatment duration and costs requires further investigation.

Strengths and limitations

The inclusion of services across diverse regions in England is a key strength of this study, offering compelling evidence of generalisability in contrast to earlier single-centre pilot studies.[11,30] The risk signalling technology was developed using historical data from a service and region that did not take part in this trial,[11] thus offering a strong test of the generalisability and predictive power of the outcome feedback model. The study was adequately powered to detect a small effect and to control for *therapist effects*. The latter feature is an important advance, confirming that the use of feedback technology improves response rates after accounting for variations in therapeutic aptitude across multiple practitioners. It should be noted that the therapist effect estimate (approximately 11%) in this study explains a considerably larger proportion of variance than the effect of feedback, so attention to the factors that characterise underperforming therapists is clearly warranted. It is, of course, plausible that some therapists may make better use of feedback than others, and future studies could investigate the personal attitudes, skills or organisational conditions that optimise adequate use of feedback.[31-33]

Some limitations should also be borne in mind when interpreting the present results. Although we included a sizeable group of therapists delivering a range of low and high intensity interventions, our study participants nevertheless volunteered to take part in the trial. We did not have information about the total size or professional characteristics of the workforce across all participating services, so we cannot assume that trial therapists are necessarily representative of the wider workforce. Furthermore, we did not have the resources to closely monitor competence in treatment delivery or in feedback utilisation. A central feature of this feedback model involves discussing risk signals with patients and clinical supervisors; however, we did not have objective data to assess the extent to which these features were adhered to. A further methodological issue relates to potential ceiling effects. Cases with high baseline severity scores (e.g., PHQ-9 ≥ 22) whose symptoms increased during treatment could not be classified as showing "reliable deterioration", which is mostly an artefact of the measurement tools and reliable change indices used in the study. It is therefore possible that the true extent of reliable deterioration rates could be underestimated. In addition, like most other feedback studies conducted to date,[8] this trial only had a short-term observation period since outcomes were assessed at the end of treatment. It is therefore unknown if the observed effects of feedback may have a durable impact on longer-term symptoms and functioning.

Conclusions

We found generalisable evidence that supplementing psychological therapy with a low-cost quality assurance system using outcome feedback technology helps to prevent deterioration in cases that are particularly prone to poor treatment outcomes.

Acknowledgements

The outcome feedback and signalling technology used in this study was developed by PCMIS at the Department of Health Sciences, University of York (http://www.pc-mis.co.uk). We thank Byron George, Gareth Percival, Colin Robson, Jim Cadwallender, Chris Jones and Andrew Bradley at PCMIS. We also thank the following colleagues who enabled us to obtain the relevant permissions and to run the study across multiple NHS Trusts: Nicole Main, Ilyas Mirza, Brenda Pimlott, Pat Mottram, James Clarke, Stephanie Ashton, Alexandra Faragher, Kathryn Simpson, Alexander Scutcher, Stephen Walker.

References

1. Lambert MJ. The efficacy and effectiveness of psychotherapy. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed.). Wiley & Sons: New York, 2013.
2. Hansen NB, Lambert MJ, Forman EM. The psychotherapy dose-response effect and its implications for treatment delivery services. *Clin Psychol-Sci Pr* 2002; **9**:329-43.
3. NHS Digital. *Psychological Therapies: Annual Report on the use of IAPT services, England, 2015-16*. Health and Social Care Information Centre, 2016. Retrieved from http://digital.nhs.uk/catalogue/PUB22110
4. Carlier IV, Meuldijk D, Van Vliet IM, Van Fenema E, Van der Wee NJ, Zitman FG. Routine outcome monitoring and feedback on physical or mental health status: evidence and theory. *J Eval Clin Pract* 2012; **18**:104-10.
5. Finch AE, Lambert MJ, Schaalje BG. Psychotherapy quality control: The statistical generation of expected recovery curves for integration into an early warning system. *Clin Psychol Psychother* 2001; **8**:231-42.
6. Shimokawa K, Lambert MJ, Smart DW. Enhancing treatment outcome of patients at risk of treatment failure: Meta-analytic and mega-analytic review of a psychotherapy quality assurance system. *J Consult Clin Psychol* 2010; **78**:298–311.
7. Castonguay LG, Barkham M, Lutz W, McAleavey AA. Practice-oriented research: approaches and applications. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed.). New York, Wiley & Sons, 2013.

8. Knaup C, Koesters M, Schoefer D, Becker T, Puschner B. Effect of feedback of treatment outcome in specialist mental healthcare: meta-analysis. *Br J Psychiatry* 2009; **195**:15-22.
9. Kendrick T, El-Gohary M, Stuart B, Gilbody S, Churchill R, Aiken L, Bhattacharya A, Gimson A, Brütt AL, de Jong K, Moore M. Routine use of patient reported outcome measures (PROMs) for improving treatment of common mental health disorders in adults. *Cochrane Libr* 2016; 7:CD011119.
10. Janse PD, De Jong K, Van Dijk MK, Hutschemaekers GJ, Verbraak MJ. Improving the efficiency of cognitive-behavioural therapy by using formal client feedback. *Psychother Res* 2017; **27**:525-38.
11. Delgadillo J, Overend K, Lucock M, Groom M, Kirby N, McMillan D, Gilbody S, Lutz W, Rubel JA, de Jong K. Improving the efficiency of psychological treatment using outcome feedback technology. *Behav Res Ther* 2017; **99**:89-97.
12. Amble I, Gude T, Stubdal S, Andersen BJ, Wampold BE. The effect of implementing the Outcome Questionnaire-45.2 feedback system in Norway: A multisite randomized clinical trial in a naturalistic setting. *Psychother Res* 2015; **25**:669-77.
13. de Jong K, Segaar J, Ingenhoven T, van Busschbach J, Timman R. Adverse Effects of Outcome Monitoring Feedback in Patients With Personality Disorders: A Randomized Controlled Trial in Day Treatment and Inpatient Settings. *J Pers Disord* 2017. DOI: 10.1521/pedi_2017_31_297
14. Lutz W, De Jong K, & Rubel J. (2015). Patient-focused and feedback research in psychotherapy: Where are we and where do we want to go? *Psychother Res* 2015; **25**: 625-632.

15. Baldwin SA, Imel ZE. Therapist Effects: Findings and Methods. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed.). Wiley & Sons: New York, 2013.
16. Spybrook J, Bloom H, Congdon R, Hill C, Liu X, Martinez A, Raudenbush S. Optimal Design Software for Multi-level and Longitudinal Research (Version 3.01) [Software], 2013. Available from www.wtgrantfoundation.org.
17. Schiefele AK, Lutz W, Barkham M, Rubel J, Böhnke J, Delgadillo J, Kopta M, Schulte D, Saxon D, Nielsen SL, Lambert MJ. Reliability of therapist effects in practice-based psychotherapy research: A guide for the planning of future studies. *Adm Policy Ment Health* 2017; **44**: 598-613.
18. Clark DM. Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: the IAPT experience. *Int Rev Psychiatry* 2011; **23**:318-27.
19. National Institute for Health and Care Excellence. *Common mental health disorders: Identification and pathways to care*. [CG123]. London, National Institute for Health and Care Excellence, 2011. Retrieved from http://www.nice.org.uk/guidance/CG123
20. Lutz W, Martinovich Z, & Howard KI. Patient profiling: An application of random coefficient regression models to depicting the response of a patient to outpatient psychotherapy. *J Consult Clin Psychol* 1999, **67**: 571-577.
21. Lutz W, Martinovich Z, Howard KI, & Leon SC. Outcome management, expected treatment response and severity adjusted provider profiling in outpatient psychotherapy. *J Clin Psychol* 2002, **58**: 1291-1304.
22. Delgadillo J, Moreea O, Lutz W. Different people respond differently to therapy: A demonstration using patient profiling and risk stratification. *Behav Res Ther* 2016; **79**:15-22.

23. Delgadillo J, Lucock M, de Jong K. Using outcome feedback in psychological therapy: A guideline for IAPT practitioners. Department of Health Sciences, University of York: York, 2015.
24. Kroenke K, Spitzer RL, & Williams JB. The PHQ-9: Validity of a brief depression severity measure. *J Gen Intern Med* 2001; **16**:606-613.
25. Richards DA, Borglin G. Implementation of psychological therapies for anxiety and depression in routine practice: two year prospective cohort study. *J Affect Disord* 2011; **133**:51-60.
26. Kroenke K, Spitzer RL, Williams JBW, Monahan PO, & Löwe B. Anxiety disorders in primary care: Prevalence, impairment, comorbidity, and detection. *Ann Intern Med* 2007; **146**:317–325.
27. Mundt JC, Marks IM, Shear MK, Greist JM. The Work and Social Adjustment Scale: a simple measure of impairment in functioning. *Br J Psychiatry* 2002; **180**:461-4.
28. Schafer JL, Olsen MK. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behav Res* 1998; **33**:545-71.
29. Raudenbush, S. W. Hierarchical linear models and experimental design. In L. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 459-496). Marcel Dekker: New York, 1993.
30. Lucock M, Halstead J, Leach C, Barkham M, Tucker S, Randal C, Middleton J, Khan W, Catlow H, Waters E, Saxon D. A mixed-method investigation of patient monitoring and enhanced feedback in routine practice: Barriers and facilitators. *Psychother Res* 2015; **25**:633-46.
31. de Jong K, van Sluis P, Nugter MA, Heiser WJ, Spinhoven P. Understanding the differential impact of outcome monitoring: Therapist variables that

moderate feedback effects in a randomized clinical trial. *Psychother Res* 2012; **22**: 464-74.

32. de Jong K, de Goede M. Why do some therapists not deal with outcome monitoring feedback? A feasibility study on the effect of regulatory focus and person–organization fit on attitude and outcome. *Psychother Res* 2015; **25**: 661-8.
33. Lutz W, Zimmermann D, Müller VN, Deisenhofer AK, Rubel JA. Randomized controlled trial to evaluate the effects of personalized prediction and adaptation tools on treatment outcome in outpatient psychotherapy: study protocol. *BMC Psychiatry* 2017; **17**: 306.

Figure 1. CONSORT diagram

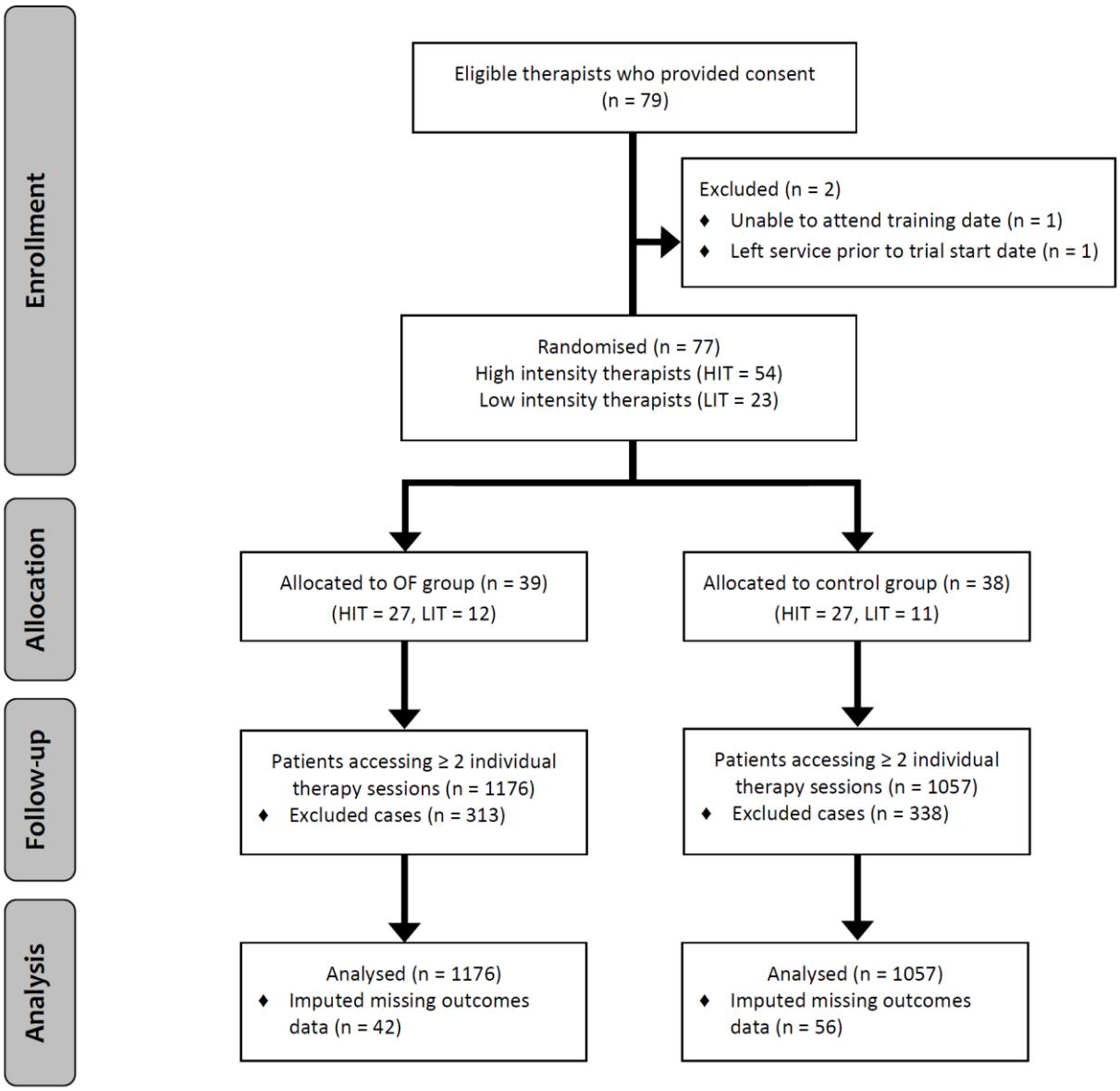

Figure 2. Differences in post-treatment depression (PHQ-9) between outcome feedback (OF) and control cases

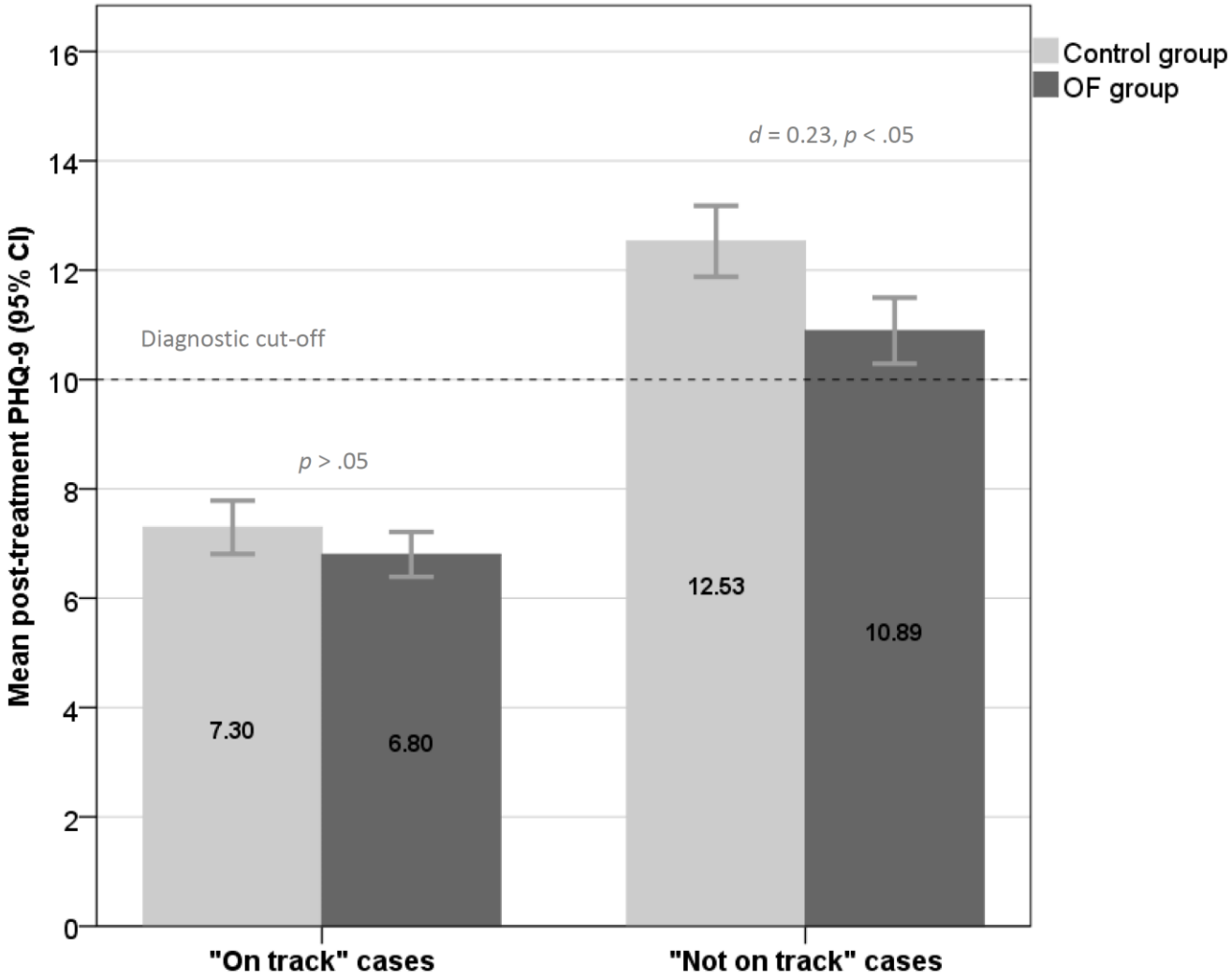

Table 1. Trial sample characteristics

	Full sample	OF group	Control group
Therapists	N = 77	N = 39	N = 38
Demographics			
Females	65 (84.4%)	30 (76.9%)	35 (92.1%)
Mean age (SD)	40.81 (11.13)	40.26 (11.29)	41.37 (11.10)
Ethnicity			
White British	65 (84.4%)	32 (82.1%)	33 (86.8%)
Other	12 (15.6%)	7 (17.9%)	5 (13.2%)
Mean years of experience (SD)	7.42 (5.79)	7.46 (5.88)	7.38 (5.77)
Treatments			
HIT	54 (70.1%)	27 (69.2%)	27 (71.1%)
LIT	23 (29.9%)	12 (30.8%)	11 (28.9%)
Patients	N = 2233	N = 1176	N = 1057
Demographics			
Females*	1465 (65.7%)	751 (63.9%)	714 (67.7%)
Mean age (SD)	39.22 (15.02)	38.40 (14.66)	40.14 (15.38)
Unemployed*	286 (18.1%)	164 (20.3%)	122 (15.7%)
Ethnicity*			
White British	1824 (88.5%)	979 (89.1%)	845 (87.8%)
Other	237 (11.5%)	120 (10.9%)	117 (12.2%)
Clinical characteristics			
Diagnosis			
Affective disorder	771 (34.5%)	413 (35.1%)	358 (33.9%)
Mixed anxiety and depression	316 (14.2%)	154 (13.1%)	162 (15.3%)
Generalized anxiety disorder	326 (14.6%)	170 (14.5%)	156 (14.8%)
Other diagnosis	820 (36.7%)	439 (37.3%)	381 (36.0%)
Baseline PHQ-9 mean (SD)	15.29 (6.20)	14.96 (5.96)	15.65 (6.43)
Baseline GAD-7 mean (SD)	13.99 (4.93)	13.82 (4.78)	14.19 (5.09)
Baseline WSAS mean (SD)	19.29 (9.40)	19.08 (9.22)	19.52 (9.57)
Mean treatment sessions (SD)	6.45 (3.67)	6.54 (3.73)	6.35 (3.60)

OF = outcome feedback; HIT = high intensity therapy; LIT = low intensity therapy; PHQ-9 = measure of depressions symptoms; GAD-7 = measure of anxiety symptoms; WSAS = work and social adjustment scale; * percentages are calculated using cases with available data, some cases with missing demographic data were excluded

Table 2. Multilevel models predicting post-treatment depression and anxiety scores

	Depression (PHQ-9) model				Anxiety (GAD-7) model			
	Fixed effects				Fixed effects			
Variable	**B**	**SE**	***p***	**95% CI**	**B**	**SE**	**p**	**95% CI**
Intercept	6.94	0.35	0.0000	6.25, 7.63	6.06	0.33	0.0000	5.42, 6.70
Sessions (Log)	-9.50	0.45	0.0000	-10.38, -8.63	-8.86	0.40	0.0000	-9.65, -8.07
Baseline severity (mc)	0.54	0.02	0.0000	0.51, 0.57	0.47	0.02	0.0000	0.43, 0.51
Group	0.19	0.49	0.69	-0.76, 1.15	0.31	0.45	0.49	-0.57, 1.20
Classification	5.64	0.30	0.0000	5.05, 6.24	5.18	0.27	0.0000	4.65, 5.71
Group * Classification	-1.03	0.41	0.012	-1.84, -0.23	-0.85	0.36	0.019	-1.56, -0.14
	Variance components (ICC = 0.107)				**Variance components (ICC = 0.114)**			
	variance	**SE**	**Z**	***p***	**variance**	**SE**	**Z**	***p***
Residual	22.04	0.66	33.30	0.0000	17.67	0.53	33.28	0.0000
Random intercept	2.63	0.59	4.45	0.0000	2.27	0.50	4.52	0.0000

Sessions: log-linear transformation for number of treatment sessions; Baseline severity (mc): mean centred values for PHQ-9 in the depression model, or GAD-7 in the anxiety model; Group: 0 = controls, 1 = Outcome Feedback cases; Classification: 0 = cases classified as "on track", 1 = cases classified as "not on track"; note that there were two symptom-specific classifications, one for PHQ-9 and one for GAD-7; Group * Classification: this interaction term is the main hypothesis test; B: regression coefficient; SE: standard error; CI: confidence intervals; ICC: intracluster correlation coefficient

Table 3. Comparison of clinical outcomes

Indicators	Full sample N = 2233		NOT subsample N = 1288	
	OF cases N = 1176	Controls N = 1057	OF cases N = 678	Controls N = 610
Clinical effectiveness				
PHQ-9 pre-treatment mean (SD)	14.41 (5.96)	14.85 (6.46)	14.47 (5.80)	15.45 (6.30)
PHQ-9 post-treatment mean (SD)	8.61 (6.60)	9.75 (7.12)	10.89 (7.17)	12.53 (7.37)
PHQ-9 Cohen's *d*	0.17		0.23	
GAD-7 pre-treatment mean (SD)	13.42 (4.85)	13.54 (5.24)	13.82 (4.77)	14.25 (5.00)
GAD-7 post-treatment mean (SD)	7.96 (5.78)	8.76 (6.12)	10.06 (6.12)	11.26 (6.37)
GAD-7 Cohen's *d*	0.13		0.19	
WSAS pre-treatment mean (SD)	19.58 (8.67)	19.88 (9.12)	20.29 (8.70)	21.03 (8.91)
WSAS post-treatment mean (SD)	12.65 (9.57)	14.11 (9.98)	15.54 (10.23)	17.72 (9.95)
WSAS Cohen's *d*	0.15		0.22	
Reliable improvement	N = 796 (67.7%)	N = 630 (59.6%)	N = 412 (60.8%)	N = 317 (52.0%)
OR (95% CI)	1.21^{ns} (0.85, 1.71)		1.32^{ns} (0.93, 1.89)	
Reliable deterioration	N = 49 (4.2%)	N = 76 (7.2%)	N = 44 (6.5%)	N = 68 (11.1%)
OR (95% CI)		1.48* (1.06, 2.07)		1.73** (1.18, 2.54)
Dropout	N = 284 (24.1%)	N = 253 (23.9%)	N = 167 (24.6%)	N = 151 (24.8%)
OR (95% CI)		1.00^{ns} (0.70, 1.43)	1.03^{ns} (0.71, 1.50)	
Classed as NOT	N = 678 (57.7%)	N = 610 (57.7%)		
OR (95% CI)	1.07^{ns} (0.86, 1.32)			

Notes: NOT = cases classified as "not on track" during therapy; PHQ-9 = depression measure; GAD-7 = anxiety measure; WSAS = work and social adjustment measure; SD = standard deviation; Cohen's d = post-treatment effect size difference between groups; OR = odds ratio, adjusting for baseline severity; * $p < 0.05$; ** $p < 0.01$; *ns* = not statistically significant

[Note: Supplementary appendices available on request to jaime.delgadillo@nhs.net]