Author's Manuscript

Note: This is a pre-print peer reviewed article. The final version will be published in a forthcoming issue of *Behaviour Research and Therapy*.

Citation: Delgadillo, J., Overend, K., Lucock, M., Groom, M., Kirby, N., McMillan, D., Gilbody, S., Lutz, W., Rubel, J.A., & de Jong, K. (in press). Improving the efficiency of psychological treatment using outcome feedback technology. *Behaviour Research and Therapy*.

DOI: 10.1016/j.brat.2017.09.011

Improving the efficiency of psychological treatment using outcome feedback technology

Jaime Delgadillo[1], Karen Overend[2], Mike Lucock[3], Martin Groom[4], Naomi Kirby[4], Dean McMillan[2], Simon Gilbody[2], Wolfgang Lutz[5], Julian A. Rubel[5] and Kim de Jong[6]

1. Clinical Psychology Unit, Department of Psychology, University of Sheffield, UK
2. Department of Health Sciences, University of York, UK
3. Centre for Applied Research in Health, University of Huddersfield, and South West Yorkshire Partnership NHS Foundation Trust, UK
4. Leeds Community Healthcare NHS Trust, UK
5. Department of Psychology, University of Trier, Germany
6. Institute of Psychology, Leiden University, The Netherlands

Declarations of interest: None.

---

* Corresponding author: jaime.delgadillo@nhs.net

Abstract

Aims: This study evaluated the impact of applying computerised outcome feedback (OF) technology in a stepped care psychological service offering low and high intensity therapies for depression and anxiety.

Methods: A group of therapists were trained to use OF based on routine outcome monitoring using depression (PHQ-9) and anxiety (GAD-7) measures. Therapists regularly reviewed expected treatment response graphs with patients and discussed cases that were "not on track" in clinical supervision. Clinical outcomes data were collected for all patients treated by this group (N = 594), six months before (controls = 349) and six months after the OF training (OF cases = 245). Symptom reductions in PHQ-9 and GAD-7 were compared between controls and OF cases using longitudinal multilevel modelling. Treatment duration and costs were compared using MANOVA. Qualitative interviews with therapists (N = 15) and patients (N = 6) were interpreted using thematic analysis.

Results: OF technology was generally acceptable and feasible to integrate in routine practice. No significant between-group differences were found in post-treatment PHQ-9 or GAD-7 measures. However, OF cases had significantly lower average duration and cost of treatment compared to controls.

Conclusions: After adopting OF into their practice, this group of therapists attained similar clinical outcomes but within a shorter space of time and at a reduced average cost per treatment episode. We conclude that OF can improve the efficiency of stepped care.

Key words: outcome feedback; stepped care; depression; anxiety; IAPT

1. Introduction

Several studies have demonstrated that monitoring patients' response to psychological treatment using standardised outcome measures can help to detect difficulties and to improve outcomes for patients (Gondek, Edbrooke-Childs, Fink, Deighton, & Wolpert, 2016). Routine outcome monitoring may be particularly important for certain patients that tend to have a poorer response to treatment (Lambert, Hansen, & Finch, 2001; Lutz, De Jong, & Rubel, 2015), referred to as 'signal cases' or cases that are 'not on track' (NOT). Lambert et al. (2003) proposed that providing timely feedback to therapists using psychometric measures to alert them about signal cases could help to improve their outcomes. Typically, outcome feedback (OF) involves entering a patient's symptom measures into a computer system that graphically displays changes from session-to-session, comparing these to clinical norms derived from hundreds of similar cases. Patients with symptoms that do not improve as suggested by these clinical norms are flagged up as NOT. A meta-analysis of controlled trials in USA concluded that NOT cases in usual psychological care were 2.3 times more likely to deteriorate by comparison to NOT cases treated by therapists that apply OF technology (Shimokawa et al., 2010). However, this meta-analysis included studies from the same research group which predominantly treated student populations, therefore raising some questions about generalizability (Davidson, Perry, & Bell, 2015). More recently, trials in European countries have replicated these findings in other clinical populations, suggesting that using OF can help to prevent deterioration in NOT cases (e.g., Amble, Gude, Stubdal, Andersen, & Wampold, 2015; De Jong et al., 2014; Hansson, Rundberg, Österling, Öjehagen, & Berglund, 2013).

Although the usefulness of outcome feedback has been demonstrated in specialist counselling and psychotherapy centres, these methods have not yet been tested in stepped care psychological services such as those linked to the IAPT (*Improving Access to Psychological Therapies*) model applied in England (Clark, 2011) and Australia (Cromarty, Drummond, Francis, Watson, & Battersby, 2016). IAPT services are particularly well placed to apply OF methods since they routinely collect standardised outcome measures at every session to monitor clinical outcomes (Clark, 2011). However, the high volume of work and time pressures typical of public healthcare settings may limit therapists' ability to consistently and meaningfully reflect on the results of outcome measures within their treatment sessions. Furthermore, research suggests that IAPT clinicians do not necessarily consider symptom measures in their decisions about treatment planning and some tend to rely on subjective beliefs and attitudes when making decisions about the treatment of non-improving patients (Delgadillo, Gellatly, & Stephenson-Bellwood, 2015). Therefore, there are plausible contextual and attitudinal barriers that may limit the effective utilization of outcome feedback in this setting.

This study presents the first application of outcome feedback technology in an IAPT stepped care context. The primary objective of the study was to evaluate the clinical impact of using OF, quantified in terms of changes in symptoms, treatment duration and cost. A secondary objective was to assess the feasibility and acceptability of discussing OF with patients in weekly therapy sessions.

2. Method

2.1. Setting, interventions and study design

This study was conducted in an IAPT stepped care service in Leeds, a large and socioeconomically diverse city in the north of England. The service offered evidence-based and protocol-driven psychological interventions for depression and anxiety problems, guided by routine session-by-session outcomes monitoring, consistent with clinical guidelines (National Institute for Health and Care Excellence, 2011). According to publicly available data for the period of the study (NHS Digital, 2016), 6410 cases referred to this service completed treatment, and this service's performance metrics for post-treatment reliable improvement (62.2%) were closely comparable to the national average (62.2%), although IAPT recovery rates (41.3%) were below the national average (46.3%).

A group of 18 psychological therapists participated in the study on a voluntary basis. The majority (N = 14) delivered high intensity cognitive behavioural therapy (CBT), 2 delivered high intensity interpersonal psychotherapy (IPT), and 2 delivered low intensity CBT. In keeping with routine practice, participating therapists could be assigned any cases on waitlist who were assessed as suitable for their treatment modality (and step of care), and no other selection of cases was applied in the study. Participating therapists routinely reviewed their patients' self-reported outcome measures (described below) at the start of every treatment session.

This was a quasi-experimental *before-and-after* study (Cook & Campbell, 1979), with concurrent economic evaluation and qualitative assessment of acceptability. Anonymised clinical records were collected for cases treated by this team (N = 594), six months before (controls = 349) and six months after they started to apply OF technology (OF cases = 245). The

dataset included all cases that came to an end of care within a 1-year study period, including completers and dropouts. Mean symptom changes, treatment duration and cost were compared statistically between controls and OF cases. Furthermore, all participating therapists (minus 3 who were members of the research team) and a consecutive sample of 6 patients (from the OF cohort) participated in semi-structured qualitative interviews conducted by the researchers, using a standard interview topic guide (available as supplementary appendix). All participants provided informed consent, their interviews were audio recorded, transcribed verbatim and analysed by a qualitative researcher. Ethical approval for the study was provided by an NHS research ethics committee (Ref: 15/NW/0675).

2.2. Outcome feedback technology

This study applied computerised OF technology that was integrated into the *Patient Case Management Information System* (PC-MIS), which is an electronic clinical record keeping system routinely used by IAPT services. The OF tool includes a graphical display of session-to-session depression and anxiety scores with overlaid clinical benchmarks, which is referred to as expected treatment response (ETR) curves. The ETR curves represent 80% confidence intervals generated using growth curve modelling following the method proposed by Finch, Lambert and Schaalje (2001). ETR curves were calculated for subgroups of cases with the same baseline severity of depression and anxiety scores, using a large clinical dataset of cases treated in IAPT (see: Delgadillo, Moreea, & Lutz, 2016). The OF tool automatically alerted therapists about NOT cases using a 'red signal', if their symptoms surpassed the 80% upper boundary of the ETR curves, and were thus progressing substantially worse than other patients.

During the control phase of the study, therapists had access to a standard version of PC-MIS, which simply plots symptom severity scores on a weekly chart, without showing ETR curves or red signals. Before the OF phase, therapists attended a 6-hour training course led by authors KdJ, ML and JD. The training covered the OF evidence base, theory and technology, and primed therapists to review and discuss ETR graphs with patients at every treatment session and to discuss NOT cases in their clinical supervision meetings.

2.3. Measures and data sources

2.3.1. Quantitative outcome measures

Patients accessing IAPT services complete two standardised outcome measures on a session-to-session basis to monitor response to treatment. The PHQ-9 is a nine-item screening tool for major depression, where each item is rated on a 0 to 3 scale, yielding a total depression severity score between 0–27 (Kroenke, Spitzer, & Williams, 2001). A cut-off ≥ 10 has been recommended to detect clinically significant depression symptoms (Kroenke, Spitzer, & Williams, 2001), and a difference of ≥6 points between assessments is indicative of reliable change (Richards & Borglin, 2011). The GAD-7 is a seven-item measure developed to screen for anxiety disorders (Spitzer, Kroenke, Williams, & Löwe, 2006). 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 (Kroenke, Spitzer, Williams, Monahan, & Löwe, 2007), and a difference of ≥5 points is indicative of reliable change (Richards & Borglin, 2011). The validity and reliability of both measures have been established

across different countries and healthcare populations (Moriarty, Gilbody, McMillan, & Manea, 2015; Plummer, Manea, Trepel, & McMillan, 2016).

2.3.2. Qualitative data sources

Standard interview topic guides (available as a supplementary appendix) were written to conduct semi-structured interviews with therapists and patients, lasting up to half an hour. Each guide had a total of 6 questions. The therapist guide aimed to explore how they used OF, their opinion about the technology, any obstacles to using OF, the influence it had on their clinical supervision and any further comments. The patient guide aimed to explore their experiences of outcome monitoring, the use of computerised technology to inform their care, and their views about how therapists should assess treatment progress.

2.3.3. Other data sources

Additional information included primary diagnosis recorded in clinical records, treatment duration (number of sessions attended), treatment completion status (versus dropout), employment status, disability (self-reported: yes/no), age, gender, ethnicity and baseline functional impairment assessed by the Work and Social Adjustment Scale (WSAS; Mundt, Mark, Shear, & Griest, 2002).

2.4. Sample characteristics

Sample characteristics are presented in Table 1 for the full sample, and each of the two study cohorts. Overall, the sample was characterised by a majority of female (63.8%) patients with a mean age of 38.69 (SD = 13.87) and from a white British background (87.8%). Approximately 40% were

unemployed and 14.1% had a self-reported disability. The most frequent diagnoses recorded in clinical records were mixed anxiety and depressive disorder (40.2%), depression (28.3%), and anxiety disorders (28.3% including post-traumatic stress disorder, generalized anxiety, panic disorder, obsessive-compulsive disorder, social phobia and other specific phobias). Mean baseline PHQ-9, GAD-7 and WSAS estimates reported in Table 1 were not significantly different between cohorts ($p > .05$).

[Table 1]

2.5. Data analysis

2.5.1. Quantitative analysis

Changes in depression (PHQ-9) and anxiety (GAD-7) scores were examined using longitudinal multilevel modelling (MLM), where session-to-session measures (level 1) were nested within cases (level 2) treated by the participating therapists. Separate models were used for each outcome measure, including random intercepts and random slopes at level 2.

MLM was performed in three steps. First, an unconditional model with no predictors was used to assess whether a linear or non-linear (quadratic, cubic, log-linear) growth trend for time (treatment sessions) provided a better fit to the data. *Goodness-of-fit* was determined by examining the AIC statistic and using log likelihood ratio tests. Preliminary tests indicated that a log-linear growth trend offered the best fit to the data (linear trend: AIC = 20081.31, -2LL = 20077.31; log-linear trend: AIC = 20029.02, -2LL = 20025.02; $x^2(1) = 52.29$, $p < 0.01$), so these settings were retained in subsequent steps. Next, we adjusted the model to control for case-mix variables (baseline severity, functional impairment, employment

status, disability, age, sessions, sessions * time interaction) given the non-randomized study design. Continuous variables were grand mean centred. Baseline severity was modelled using a factor score which combined all items from PHQ-9 and GAD-7 to reduce multicollinearity. In the third step, we entered a binary 'cohort' variable as a predictor (controls vs. OF cases), as well as a cohort * time interaction term.

Using the diagnostic cut-offs and reliable change indices for each outcome measure (described above), we applied Jacobson and Truax (1991) criteria to report the proportions of cases with reliable and clinically significant improvement (RCSI) and reliable deterioration during their stepped care treatment episode. RCSI rates, treatment dropout rates and total cases classed as NOT were compared between cohorts using case-mix adjusted logistic regression. Supplementary outcome metrics commonly used in IAPT services were also estimated and described, including reliable improvement rates (Jacobson & Truax, 1991) for each measure and IAPT recovery rates (Clark et al., 2009).

Every case was assigned a treatment cost by multiplying the average hourly rate for each professional group (NHS pay grades: band 5, 6 and 7, including organisational overheads) by the total number of contact hours recorded in clinical records during the entire stepped care pathway. The mean number of treatment contacts and average direct treatment costs were compared between cohorts using MANOVA, which allows the values of multiple dependent scale variables to be modelled in a single analysis based on their relationships to predictors (case-mix variables).

Each of the above analyses were applied in the full sample (N = 594), and repeated as secondary analyses in the subsamples of cases that were classed as NOT (N = 318) and OT (N = 276).

2.5.2. Qualitative analysis

Qualitative interview transcripts from 15 therapists and 6 patients were analysed together by a primary reviewer following the six phases of thematic analysis described by Braun and Clarke (2006). Stage one involved familiarisation with all transcripts. Stage two involved 'open coding' through a line-by-line inspection of transcripts. Stage three involved clustering codes into potential themes through constant comparison within and across transcripts. Stage four involved generating a thematic map. Stage five aimed to refine the themes into a coherent narrative structure. Finally, stage six involved the selection of representative data extracts to produce a descriptive account. A secondary reviewer independently analysed a subset of transcripts, and notes were compared between reviewers to refine the thematic map using a constant comparison and peer review approach (Angen, 2000).

3. Results

3.1. Quantitative data on clinical impact

Fixed effects of the fully adjusted MLM analysis are shown in Table 2. The cohort * time interaction term represents the main between-group comparison in symptom changes across time. This was not statistically significant in PHQ-9 (B = 0.80, SE = 0.78, p = 0.30) or GAD-7 (B = 0.80, SE = 0.71, p = 0.26) models applied in the full sample, nor in the subsamples of NOT (shown in Table 2) or OT cases.

[Table 2]

Case-mix adjusted logistic regressions (Table 3) indicated that RCSI rates were not significantly different between controls and OF cases; PHQ-9: 41.8% vs. 38.2%, OR = 1.01, 95% CI [0.67, 1.52], p = 0.96; GAD-7: 31.6% vs. 29.5%, OR = 1.20, 95% CI [0.81, 1.78], p = 0.36. The overall numbers of cases with reliable deterioration were too small to compare statistically; with 19 cases in the control cohort and 6 cases in the OF cohort. No significant differences in dropout rates were found between controls (27.3%) and OF cases (27.6%); B = -0.03, SE = 0.21, p = 0.90, OR = 0.98, 95% CI [0.65, 1.46]. Secondary analyses in NOT and OT samples across various outcome metrics presented in tables 3 and 4 yielded the same results as above. However, cases in the control cohort were significantly more likely to be classed as NOT by comparison to OF cases; B = 0.81, SE = 0.18, p = <0.001, OR = 2.25, 95% CI [1.58, 3.20].

[Tables 3 and 4]

MANOVA results indicated that the mean number of treatment contacts in the control cohort (adjusted mean = 10.25, SE = 0.45) was significantly higher than the OF cohort (adjusted mean = 6.59, SE = 0.51); B = 3.66, SE = 0.55, p < 0.001; SMD = 3.66, 95% CI [2.58, 4.74]. This is illustrated in Figure 1 which displays trajectories of change (growth curves) in depression symptoms over time for controls and OF cases. The OF group has a shorter curve since the range of treatment length was between 1 to 20 sessions, whereas the control group had a longer range between 1 and 36 sessions. Although the confidence intervals (dashed lines) for both curves overlap, the figure shows a trend for lower-level symptoms in the OF group, which is plausibly explained by the significantly lower percentage of cases

classified as NOT in that group. As shown in Figure 2, the average cost of treatment was significantly higher for controls (adjusted mean = £246.43, SE = 13.24) by comparison to OF cases (adjusted mean = £148.90, SE = 14.97); B = 97.54, SE = 16.12, $p < 0.001$, SMD = £97.54, 95% CI [£65.88, £129.20]. The average treatment duration and cost estimates were also significantly higher for the control cohort in the samples of NOT (SMD: sessions = 2.77 [1.23, 4.32]; cost = £75.65 [£29.14, £122.15]) and OT cases (SMD: sessions = 3.02 [1.60, 4.45]; cost = £74.56 [£34.60, £114.52]).

[Figures 1 and 2]

3.2. Qualitative data on feasibility and acceptability

Three overarching themes emerged through constant comparison of qualitative interview transcripts; these are described below with reference to participant quotes (where T = therapist, P = patient).

Theme 1: Implementing outcome feedback (OF)

Most therapists discussed the rationale behind OF at the first or second therapy session, and they tended to review Expected Treatment Response (ETR) charts at the start of every session. This practice was corroborated by the majority of patients [5 of 6] who reported that OF was used on a weekly basis.

> *"every week the item on the agenda was always checking in with my current mood on that day so we sort of –you know– looked at how I was feeling, literally within five minutes of the session."* P6, 9-11.

This process prompted therapists to raise problem solving conversations related to NOT signals, but also to motivate and reinforce positive change when patients' symptoms were on track. A few therapists reported reviewing ETR charts with patients at less frequent intervals, such as every 4 to 6 weeks.

Therapists reported that ETR charts alerted them to unnoticed difficulties and enabled them to review their treatment plan in collaboration with patients.

> *"There were a couple of occasions where it highlighted to both of us really that the treatment we were doing –although it was useful for them– it wasn't as effective as it could be, so it meant that we could change very quickly what we were doing."* T12, 49-51.

The ETR charts were seen as a helpful tool to support and sometimes to correct clinical judgement.

> *"It's been a useful tool in getting me to think… I can sometimes blindly continue thinking 'this is going to work, we'll get some effect', and it's allowed me to say actually say no – we need to take stock"* T3, 116-123.

One patient indicated that OF gave their therapist an insight into what was working and highlighted which aspects of therapy to focus on.

> "I*t helped [the therapist] know where we were going with our sessions and it helped me understand what was working and what wasn't"* P1, 30-33.

Therapists also reported using ETR charts to reflect about treatment progress and to inform treatment planning within clinical supervision meetings. Some therapists stated that this method prompted them to discuss some cases in clinical supervision earlier than they would normally do, and sometimes prompted decisions to 'step patients up' to more intensive treatments. A few therapists also stated that they were using ETR charts to inform their professional development plans as part of clinical supervision meetings.

> *"We pull up the graph, go into the client's record and see what their progress has been like. And from then my supervisor will ask me if there have been any potential barriers, anything that got in the way of the client progressing on track as we would hope to expect, and it helps me to be able to reflect"* T12 160-168.

Theme 2: Experiences and acceptability

Therapists found the ETR charts and system easy to use, compatible with usual outcome monitoring in IAPT, and they were able to integrate it within sessions without much difficulty.

> *"It's really handy, it's simple that's the good thing. As a user you can look at it and spot things quickly"* T2, 65-69.

Most patients also found it interesting and useful.

> *"It's definitely a useful tool because sometimes you don't realise you've made progress, but if you've got something on screen showing you what your scores were and what they are, it quantifies your progress"* P4, 74-78.

One exception was reported by a patient accessing low intensity CBT (30 minute sessions) who felt it took up too much time.

> *"I wouldn't have minded ten minutes of an hour, but ten minutes of half an hour is... it cuts your time down, my time down"* P3, 53-56.

Many therapists described how OF helped to involve and engage patients, thereby enhancing collaboration.

> *"For me the biggest thing is about the increase in the collaboration with yourself and the patient, also creating a transparency of what these measures are for, I found that helpful, and it boosts the relationship"* T13; 149-150.

Some patients also suggested that reviewing ETR charts prompted therapists to enable them to reflect and gain insights about their problems.

> *"My therapist was really good at picking out and getting me to talk about stuff that –um you know– that came to mind –you know– that I hadn't realised before I did it"* P2, 30-31.

Therapists at times felt that having conversations about NOT signals could be daunting; but at other times they found that the ETR system gave them confidence to raise difficult issues.

> *"It's just given me a bit more confidence that it's okay to have those difficult conversations with people"* T15, 137-142.

It was also reported that reviewing ETR graphs can boost therapists' confidence and provide reassurance about their practice when cases were on track.

Theme 3: Challenges and solutions

Therapists described a series of challenges that they encountered as they started to implement OF in routine practice. Some challenges were of a technical nature (i.e., lack of computer in clinic room), which were possible to resolve by printing ETR graphs in advance of sessions. Other challenges related to explaining to patients how ETR boundaries were calculated in order to clarify the rationale for OF.

> *"The only difficulty I found is trying to explain how we come up with the status. What I've been saying to people is [that] we use data for people who started at the same score as you"* T3, 62-64.

Some therapists raised examples of patients who did not like filling in questionnaires, and others who did not complete them accurately. A common theme in these discussions was the importance of how ETR charts are explained to patients in order to foster collaboration with the OF method. Therapists reflected on the importance of using lay terminology to explain the rationale for OF, and using non-threatening language when discussing NOT signals.

> *"I think the most difficult thing was how we describe it to patients without putting too much emphasis on the expected treatment outcomes. [...] Just trying to say 'this is the average' rather than 'this is what we expect' because I don't want*

people to feel they're not meeting the expectations" T12, 65-69.

4. Discussion

4.1. Main findings

This study presents the first comprehensive evaluation of outcome feedback technology applied in a stepped care psychological treatment setting. The results indicated that this technology was feasible to adopt in routine care, was minimally burdensome, and was generally seen by therapists and patients as a useful aid to decision-making and clinical supervision processes.

Qualitative interviews with therapists revealed that the outcome feedback signalling technology influenced their interpretation and use of routine outcomes data in a number of ways: they openly discussed outcomes data with patients more consistently; they tended to take notice of obstacles to improvement much sooner than usual; they prioritised NOT cases in clinical supervision meetings; they were more open to the possibility that their clinical impressions may be incorrect or in need of revision and consequently they were also more open to reconsider their treatment plan. Therapists also stated that they felt more confident in explaining the rationale of outcome monitoring to their patients and in discussing and addressing potential problems, which they felt enhanced collaboration. Furthermore, although quantitative data on the frequency of these outcome monitoring discussions were not collected, the qualitative interviews reflected a remarkable consistency in therapists' and patients' accounts of the regular and collaborative use of outcome feedback. Some individual differences were also apparent in therapists' way of explaining ETR charts

and the frequency with which they discussed these with patients. Overall, and considering that this group of therapists had years of experience collecting and reviewing outcome measures, the OF signalling technology made outcome monitoring a more salient, informative and collaborative aspect of the treatment process.

The integration of routine outcome monitoring in IAPT stepped care services (Clark, 2011) is likely to have supported a culture and infrastructure that is open to innovations like ETR technology. This technical and cultural readiness may explain the ease of adoption reflected in this study, which stands in contrast to previous efforts to implement feedback methods in psychotherapy services that are less accustomed to routine outcome monitoring (e.g., Gleacher et al., 2016; Lucock et al., 2015).

Quantitative analyses revealed that applying OF technology yielded similar outcomes to usual care, but within fewer sessions and at lower cost, considerably enhancing the efficiency of treatment. Furthermore, control cases were twice as likely to be classed as NOT by comparison to OF cases (OR = 2.25). The standardised mean difference (cost saving) for an average treatment was approximately £97.54 [£65.88, £129.20]. Taking the conservative lower bound of the confidence interval (£65.88) and multiplying this by the total of 245 OF cases equates to an estimated cost saving of £16,140.60 in the treatment of that cohort of patients within 6 months. This converges with a recent study that also found OF to yield similar outcomes more efficiently in CBT interventions (Janse, De Jong, Van Dijk, Hutschemaekers, & Verbraak, 2017). This replication of findings indicates that using OF technology can enhance the efficiency of psychological care in settings where protocol-driven CBT is a predominant treatment model.

Current wisdom in the field suggests that OF is specifically helpful for a subset of signal cases that are classified as NOT during therapy (Carlier et al., 2012; Castonguay, Barkham, Lutz, & McAleavey, 2013; Knaup, Koesters, Schoefer, Becker, & Puschner, 2009; Shimokawa, Lambert, & Smart, 2010). In fact, some reviews that examine controlled trials of feedback in full samples (rather than subsamples of NOT cases) fail to detect the effects that are typically observed in these studies (Kendrick et al., 2016). Contrary to findings reported in most reviews, we found no evidence of differential effects of feedback on the clinical outcomes of NOT cases. It is possible that the standard outcome monitoring technology that supported the treatment of control cases could already be working as a useful feedback tool, possibly explaining the lack of differences in clinical outcomes between-groups. It is also apparent that using OF technology considerably reduces the chances of being classed as NOT during treatment; hence another explanation is that the few cases classed as NOT in the OF cohort could be those which are generally unresponsive to psychological interventions.

The confluence of qualitative and quantitative evidence in this study suggests that feedback in this setting may work by alerting therapists to identify and to resolve obstacles sooner, thus accelerating the recovery process for cases that are amenable to therapeutic improvement, as well as by providing an earlier signal to 'step up' cases that are clearly not responding to treatment. An alternative explanation could be that the training and emphasis that the study placed on regular review of outcome measures in clinical sessions and supervision meetings may account for the observed effect, regardless of the specific effect of NOT signals. It should be noted that the study design did not enable us to isolate the specific influence

of NOT signals and to disentangle this from more conventional aspects of diligent and collaborative outcome monitoring.

4.2. Strengths and limitations

The large sample size of cases treated in a routine care context and *intention-to-treat* approach to analyses enhance the external validity of the study. The *before-and-after* design also enabled us to minimise confounding due to therapist effects, since each therapist was his/her own control. However, the lack of random allocation and the use of historical controls raise some threats to internal validity. We cannot rule out the possibility that unmeasured external influences (i.e., policy, managerial or practice changes) could have influenced the length of interventions. An important caveat to our findings is that this self-selected group of therapists may be particularly motivated to apply OF in a way that may not be representative of the wider IAPT programme workforce. Furthermore, we only interviewed a small number of patients, who may not be representative of the wider clinical population. A larger-scale, multi-service randomised controlled trial is necessary to gain a more rigorous and generalizable view about the impact of feedback and the degree to which IAPT services are 'ready' to adopt these methods in routine care. Future studies could also assess outcomes using different outcome measures to those used as part of the feedback process to assess the extent to which tracking generic anxiety measures like GAD-7 impacts on outcomes assessed using disorder-specific measures for conditions like post-traumatic stress disorder, obsessive compulsive disorder, panic disorder, etc.

A further limitation is that this study did not collect follow-up data after the end of the acute phase of treatment, so it is unclear whether the apparent benefits of treatment are durable in the long run. We do not know,

from the available data, if patients in the OF condition may have been discharged from therapy shortly after attaining remission of symptoms, or if some treatment sessions after initial remission were appropriately devoted to relapse prevention in order to maintain longer-term gains. The growth curve corresponding to the OF sample in Figure 1 clearly decelerates (flattens) between sessions 15 and 20, which would indicate a trend of ongoing treatment sessions after average remission of symptoms. This suggests that treatment continued for some time after initial remission of symptoms in the OF group (possibly devoted to consolidating gains and relapse prevention, although we did not have data on session content); whereas treatment continued for considerably longer in control group cases that were apparently unresponsive to treatment. An important direction for future research is to investigate longer-term remission and relapse rates after brief psychological interventions assisted by outcome feedback technology, to ensure that apparent gains in efficiency do not come at the expense of longer-term relapse.

4.3. Conclusions

Outcome feedback technology was feasible to implement in routine practice, it was generally acceptable to therapists and patients, and was associated with improved efficiency and reduced costs of stepped care psychological treatment. The gains in efficiency could enable therapists to invest additional time and effort in relapse prevention to maximise long-term recovery for those who respond to treatment. Conversely, cases that are clearly not responding to treatment can be detected using OF and appropriately stepped up sooner, as per stepped care guidelines.

Acknowledgements

The Leeds Outcome Feedback Study (NHS REC Reference: 15/NW/0675) was supported by research capability funding awarded by Leeds Community Healthcare NHS Trust. 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, Alexander Teahan, Jan Thomson, Simon Day, Anne Briggs, Jan Lewis, Abigail Coe, Sarah West, Caroline Lloyd, Angela Lennon, David Shanley, Jill Hardwick, Matthew Garner, Barbara Howard, Teresa Bolton, Toby Chelms, Kate Crompton and Jacks Hillaby.

References

Amble, I., Gude, T., Stubdal, S., Andersen, B.J., & Wampold, B.E. (2015). The effect of implementing the Outcome Questionnaire-45.2 feedback system in Norway: A multisite randomized clinical trial in a naturalistic setting. *Psychotherapy Research*, 25(6), 669-677.

Angen, M.J. (2000). Evaluating interpretive inquiry: reviewing the validity debate and opening the dialogue. *Qualitative Health Research, 10*(3), 378–95.

Braun, V., and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*, 77–101.

Carlier, I. V. E., Meuldijk, D., Van Vliet, I. M., Van Fenema, E., Van der Wee, N. J. A., & Zitman, F. G. (2012). Routine outcome monitoring and feedback on physical or mental health status: evidence and theory. *Journal of Evaluation in Clinical Practice, 18*, 104-110.

Castonguay, L. G., Barkham, M., Lutz, W., & McAleavey, A. A. (2013). 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.

Clark, D. M., Layard, R., Smithies, R., Richards, D. A., Suckling, R., & Wright, B. (2009). Improving access to psychological therapy: Initial evaluation of two UK demonstration sites. *Behaviour Research and Therapy, 47* (11), 910-920.

Clark D. M. (2011). Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: The IAPT experience. *International Review of Psychiatry, 23*, 318–327.

Cook, T.D., Campbell, D.T. (1979). *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand McNally.

Cromarty, P., Drummond, A., Francis, T., Watson, J., & Battersby, M. (2016). NewAccess for depression and anxiety: adapting the UK Improving Access to Psychological Therapies Program across Australia. Australasian Psychiatry, doi: 1039856216641310.

Davidson, K., Perry, A., & Bell, L. (2015). Would continuous feedback of patient's clinical outcomes to practitioners improve NHS psychological therapy services? Critical analysis and assessment of quality of existing studies. *Psychology and Psychotherapy: Theory, Research and Practice, 88*(1), 21-37.

Delgadillo, J., Gellatly, J., & Stephenson-Bellwood, S. (2015). Decision Making in Stepped Care: How Do Therapists Decide Whether to Prolong Treatment or Not? *Behavioural and Cognitive Psychotherapy*, 43(3), 328-41.

Delgadillo, J., Moreea, O., & Lutz, W. (2016). Different people respond differently to therapy: A demonstration using patient profiling and risk stratification. *Behaviour Research and Therapy*, 79, 15-22.

De Jong, K., Timman, R., Hakkaart-Van Roijen, L., Vermeulen, P., Kooiman, K., Passchier, J., & Van Busschbach, J. (2014). The effect of outcome monitoring feedback to clinicians and patients in short and long-term psychotherapy: A randomized controlled trial. Psychotherapy Research, 24(6), 629-639.

Finch, A. E., Lambert, M. J., & Schaalje, B. G. (2001). Psychotherapy quality control: the statistical generation of expected recovery curves for integration into an early warning system. *Clinical Psychology and Psychotherapy*, 8(4), 231-242.

Gleacher, A.A., Olin, S.S., Nadeem, E., Pollock, M., Ringle, V., Bickman, L., Douglas, S. and Hoagwood, K. (2016). Implementing a measurement feedback system in community mental health clinics: A case study of multilevel barriers and facilitators. *Administration and Policy in Mental Health and Mental Health Services Research*, *43*(3), 426-440.

Gondek, D., Edbrooke-Childs, J., Fink, E., Deighton, J., & Wolpert, M. (2016). Feedback from outcome measures and treatment effectiveness, treatment efficiency, and collaborative practice: A systematic review. *Administration and Policy in Mental Health and Mental Health Services Research*, 43(3), 325-343.

Hansson, H., Rundberg, J., Österling, A., Öjehagen, A., & Berglund, M. (2013). Intervention with feedback using Outcome Questionnaire 45 (OQ-45) in a Swedish psychiatric outpatient population. A randomized controlled trial. *Nordic Journal of Psychiatry*, 67(4), 274-281.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 59(1), 12-19.

Janse, P. D., De Jong, K., Van Dijk, M. K., Hutschemaekers, G. J., & Verbraak, M. J. (2017). Improving the efficiency of cognitive-behavioural therapy by using formal client feedback. *Psychotherapy Research*, 27(5), 525-538.

Kendrick, T., El-Gohary, M., Stuart, B., Gilbody, S., Churchill, R., Aiken, L., Bhattacharya, A., Gimson, A., Brütt, A.L., de Jong, K. and Moore, M. (2016). Routine use of patient reported outcome measures (PROMs) for improving treatment of common mental health disorders in adults. The Cochrane Library, Issue 7. Art. No.: CD011119. doi: 10.1002/14651858.CD011119.pub2.

Knaup, C., Koesters, M., Schoefer, D., Becker, T., & Puschner, B. (2009). Effect of feedback of treatment outcome in specialist mental healthcare: meta-analysis. *British Journal of Psychiatry, 195*, 15-22

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9), 606–613.

Kroenke, K., Spitzer, R. L., Williams, J. B. W., Monahan, P. O., & Löwe, B. (2007). Anxiety disorders in primary care: Prevalence, impairment, comorbidity, and detection. *Annals of Internal Medicine*, 146(5), 317–325.

Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology*, 69(2), 159.

Lambert, M. J., Whipple, J. L., Hawkins, E. J., Vermeersch, D. A., Nielsen, S. L., & Smart, D. W. (2003). Is it time for clinicians to routinely track patient outcome? A meta-analysis. *Clinical Psychology: Science and Practice*, 10(3), 288–301.

Lucock, M., Halstead, J., Leach, C., Barkham, M., Tucker, S., Randal, C., Middleton, J., Khan, W., Catlow, H., Waters, E. and Saxon, D. (2015). A mixed-method investigation of patient monitoring and enhanced feedback in routine practice: Barriers and facilitators. *Psychotherapy Research*, *25*(6), 633-646.

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? *Psychotherapy Research*, *25*(6), 625-632.

Moriarty, A. S., Gilbody, S., McMillan, D., & Manea, L. (2015). Screening and case finding for major depressive disorder using the Patient Health Questionnaire (PHQ-9): a meta-analysis. *General Hospital Psychiatry, 37*(6), 567-576.

Mundt, J. C., Mark, I. M., Shear, M. K., Griest, J. M. (2002). The Work and Social Adjustment Scale: a simple measure of impairment in functioning. *British Journal of Psychiatry*, 180(5), 461-464.

National Institute for Health and Care Excellence. (2011). *Common mental health disorders: Identification and pathways to care* [CG123]. London: National Institute for Health and Care Excellence. Retrieved from http://www.nice.org.uk/guidance/CG123.

NHS Digital. (2016). *Improving Access to Psychological Therapies (IAPT) dataset reports*. Leeds, UK: Health and Social Care Information Centre, 2016. Retrieved from http://content.digital.nhs.uk/iaptreports

Plummer, F., Manea, L., Trepel, D., & McMillan, D. (2016). Screening for anxiety disorders with the GAD-7 and GAD-2: a systematic review and diagnostic metaanalysis. *General Hospital Psychiatry, 39*, 24-31.

Richards, D. A., & Borglin, G. (2011). Implementation of psychological therapies for anxiety and depression in routine practice: Two year prospective cohort study. *Journal of Affective Disorders*, *133*, 51-60.

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

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

Table 1. Sample characteristics

	Full sample (N = 594)	**Cohort 1: controls** (N = 349)	**Cohort 2: OF cases** (N = 245)
Demographics			
Mean age (SD)	38.69 (13.78)	38.49 (13.65)	38.96 (13.99)
Females (%)	63.8	62.5	65.7
White British (%)	87.8	87.9	87.6
Unemployed (%)	40.0	39.8	40.2
Disabled (%)	14.1	13.1	15.5
Primary diagnosis			
Depression (%)	28.3	26.1	31.2
Anxiety disorder (%)	28.3	31.7	23.8
*Mixed Anx & Dep (%)	40.2	39.2	41.5
Eating disorder (%)	2.1	2.6	1.5
Somatoform disorder (%)	1.1	0.4	2.0
Baseline severity of symptoms and functioning			
Mean PHQ-9 (SD)	14.42 (6.33)	14.59 (6.26)	14.17 (6.42)
Mean GAD-7 (SD)	12.87 (5.40)	13.17 (5.35)	12.43 (5.46)
Mean WSAS (SD)	20.02 (9.17)	20.52 (9.07)	19.30 (9.28)

Notes: * mixed anxiety and depressive disorder; OF = outcome feedback; SD = standard deviation; PHQ-9 = depression severity; GAD-7 = anxiety severity; WSAS = functional impairment

Table 2. Longitudinal multilevel modelling comparing outcome changes between controls and OF cases

		Fixed effects					
		Full sample (N = 594)			NOT sample (N = 318)		
Outcome	Variable	B	SE	p	B	SE	p
PHQ-9	Intercept	17.70	0.68	<0.001	18.82	0.78	<0.001
	Time (Log)	-7.34	0.66	<0.001	-6.86	0.87	<0.001
	Age	-0.02	0.01	0.26	0.00	0.02	0.81
	Factor score	2.93	0.24	<0.001	3.12	0.29	<0.001
	WSAS	0.15	0.03	<0.001	0.13	0.03	<0.001
	Employed (vs. unemployed)	-0.46	0.40	0.25	-0.49	0.45	0.28
	Not disabled (vs. disabled)	-1.90	0.61	<0.01	-1.19	0.66	0.07
	Cohort(1) (vs. 2)	0.07	0.50	0.89	-0.80	0.62	0.20
	Cohort * Time	**0.80**	**0.78**	**0.30**	**0.82**	**1.00**	**0.41**
	Sessions	0.06	0.04	0.14	0.01	0.05	0.84
	Sessions * Time	0.06	0.06	0.35	0.02	0.08	0.75
GAD-7	Intercept	15.12	0.61	<0.01	16.04	0.69	<0.001
	Time (Log-linear)	-6.65	0.60	<0.01	-6.25	0.81	<0.001
	Age	-0.03	0.01	0.03	-0.01	0.01	0.49
	Factor score	2.59	0.22	<0.01	2.66	0.25	<0.001
	WSAS	0.08	0.02	0.00	0.05	0.03	0.04
	Employed (vs. unemployed)	-0.35	0.36	0.34	-0.38	0.39	0.33
	Not disabled (vs. disabled)	-1.08	0.55	0.05	-0.11	0.57	0.85
	Cohort(1) (vs. 2)	0.21	0.45	0.64	-0.23	0.55	0.67
	Cohort * Time	**0.80**	**0.71**	**0.26**	**0.44**	**0.93**	**0.63**
	Sessions	0.11	0.04	<0.01	0.01	0.04	0.84
	Sessions * Time	0.02	0.06	0.79	0.02	0.07	0.77

Notes: NOT = not on track; SE = standard error; PHQ-9 = depression severity; GAD-7 = anxiety severity; WSAS = functional impairment; Cohort 1 = controls; Cohort 2 = OF cases; all continuous measures were grand mean centred; **main hypothesis test in bold text**

Table 3. Logistic regressions comparing outcomes between controls and OF cases

		Full sample (N = 594)				NOT sample (N = 318)			
Outcome	Variable	B	SE	p	OR	B	SE	p	OR
PHQ-9 RCSI									
	Age	0.01	0.01	0.49	1.01	0.01	0.01	0.50	1.01
	Factor score	-0.25	0.15	0.10	0.78	-0.16	0.19	0.40	0.85
	WSAS	-0.03	0.01	0.03	0.97	-0.04	0.02	0.05	0.96
	Disabled (vs. not disabled)	-0.72	0.34	0.03	0.49	-0.49	0.40	0.23	0.61
	Unemployed (vs. employed)	-0.39	0.21	0.06	0.68	-0.36	0.27	0.19	0.70
	Cohort 2 (vs. 1)	**0.01**	**0.21**	**0.96**	**1.01**	**-0.06**	**0.29**	**0.84**	**0.95**
	Constant	0.46	0.44	0.30	1.58	0.47	0.56	0.40	1.60
GAD-7 RCSI									
	Age	0.01	0.01	0.25	1.01	0.01	0.01	0.43	1.01
	Factor score	-0.14	0.14	0.33	0.87	-0.14	0.18	0.43	0.87
	WSAS	-0.05	0.01	0.00	0.95	-0.06	0.02	0.00	0.95
	Disabled (vs. not disabled)	-0.61	0.32	0.06	0.54	-0.49	0.40	0.22	0.61
	Unemployed (vs. employed)	-0.46	0.21	0.03	0.63	-0.52	0.27	0.05	0.59
	Cohort 2 (vs. 1)	**0.19**	**0.20**	**0.36**	**1.20**	**0.28**	**0.28**	**0.31**	**1.32**
	Constant	0.74	0.43	0.08	2.09	0.77	0.54	0.15	2.15
Dropout									
	Age	-0.02	0.01	<0.001	0.98	-0.02	0.01	0.11	0.98
	Factor score	-0.04	0.13	0.78	0.97	0.14	0.18	0.43	1.15
	WSAS	0.02	0.01	0.12	1.02	0.02	0.02	0.27	1.02
	Disabled (vs. not disabled)	-0.03	0.31	0.92	0.97	0.33	0.38	0.39	1.39
	Unemployed (vs. employed)	0.42	0.20	0.04	1.52	0.31	0.27	0.25	1.36
	Cohort 2 (vs. 1)	**-0.03**	**0.21**	**0.90**	**0.98**	**0.02**	**0.29**	**0.94**	**1.02**
	Constant	-0.27	0.20	0.18	0.76	-1.08	0.58	0.06	0.34
% Cases classified as NOT									
	Age	0.00	0.01	0.59	1.00				
	Factor score	0.06	0.11	0.58	1.06				
	WSAS	0.02	0.01	0.07	1.02				
	Disabled (vs. not disabled)	0.04	0.27	0.89	1.04				
	Unemployed (vs. employed)	0.03	0.18	0.88	1.03				
	Cohort 1 (vs. 2)	**0.81**	**0.18**	**<0.001**	**2.25**				
	Constant	0.20	0.39	0.60	1.23				

Notes: NOT = not on track; SE = standard error; OR = odds ratio; PHQ-9 = depression severity; GAD-7 = anxiety severity; WSAS = functional impairment; Cohort 1 = controls; Cohort 2 = OF cases; RCSI = reliable and clinically significant improvement; **main hypothesis test in bold text**

Table 4. Summary of clinical outcomes in controls and OF cases

	Full sample (N = 594)		NOT sample (N = 318)	
Outcomes	Cohort 1 (controls)	Cohort 1 (OF cases)	Cohort 1 (controls)	Cohort 1 (OF cases)
Pre- Mean PHQ-9 (SD)	14.59 (6.26)	14.17 (6.42)	15.31 (5.86)	16.63 (5.86)
Post- Mean PHQ-9 (SD)	10.84 (7.48)	11.05 (7.03)	11.13 (7.38)	12.60 (7.57)
Pre- Mean GAD-7 (SD)	13.17 (5.35)	12.43 (5.46)	14.31 (4.72)	14.68 (5.09)
Post- Mean GAD-7 (SD	9.43 (6.13)	9.56 (5.95)	9.85 (5.88)	10.86 (6.60)
PHQ-9 RCSI (%)	41.8	38.2	42.1	34.4
PHQ-9 RI (%)	49.9	47.6	48.6	43.1
PHQ-9 RD (%)	5.6	2.6	6.5	4.9
GAD-7 RCSI (%)	31.6	29.5	39.2	38.4
GAD-7 RI (%)	50.1	48.5	50.5	47.1
GAD-7 RD (%)	4.4	2.6	5.1	3.9
IAPT recovery (%)	50.8	44.9	50.3	42.0

Notes: SD = standard deviation of the mean; NOT = not on track; PHQ-9 = depression severity; GAD-7 = anxiety severity; RCSI = reliable and clinically significant improvement; RI = reliable improvement; RD = reliable deterioration; IAPT recovery = cases where at least one measure (PHQ-9 or GAD-7) was in the clinical range at baseline and where both measures were below the clinical cut-offs post-treatment

Figure 1.
Trajectories of change in depression symptoms comparing controls and Outcome Feedback (OF) cases

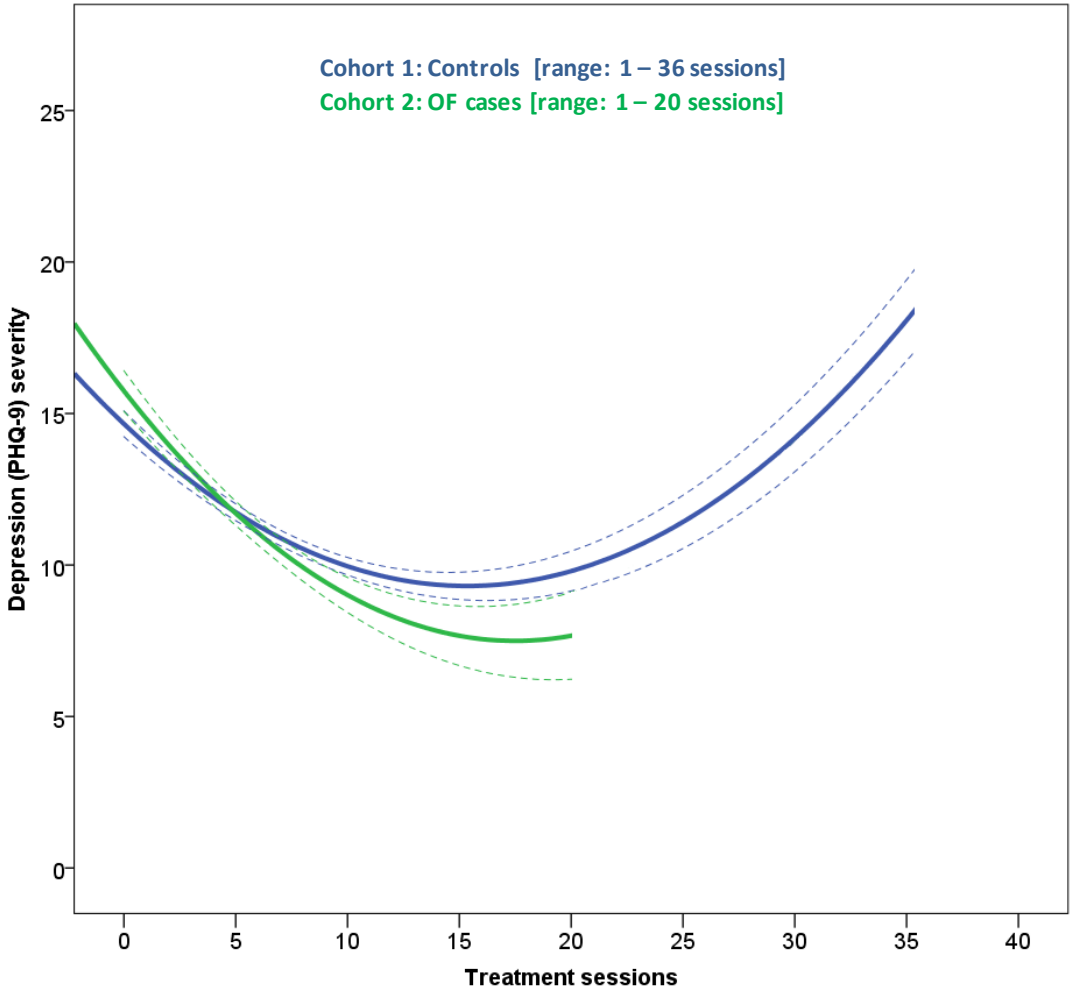

Figure 2.
Standardised mean difference (SMD) in treatment costs between controls and OF cases