Cochrane Library

Cochrane Database of Systematic Reviews

Routine use of patient reported outcome measures (PROMs) for improving treatment of common mental health disorders in adults (Review)

Kendrick T, El-Gohary M, Stuart B, Gilbody S, Churchill R, Aiken L, Bhattacharya A, Gimson A, Brütt AL, de Jong K, Moore M

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 Database of Systematic Reviews 2016, Issue 7. Art. No.: CD011119.
DOI: 10.1002/14651858.CD011119.pub2.

www.cochranelibrary.com

WILEY

TABLE OF CONTENTS

[Intervention Review]

Routine use of patient reported outcome measures (PROMs) for improving treatment of common mental health disorders in adults

Tony Kendrick[1], Magdy El-Gohary[1], Beth Stuart[1], Simon Gilbody[2], Rachel Churchill[3], Laura Aiken[4], Abhishek Bhattacharya[4], Amy Gimson[5], Anna L Brütt[6], Kim de Jong[7], Michael Moore[1]

[1]Primary Care and Population Sciences, Faculty of Medicine, Aldermoor Health Centre, University of Southampton, Southampton, UK. [2]Department of Health Sciences, University of York, York, UK. [3]Centre for Academic Mental Health, School of Social and Community Medicine, University of Bristol, Bristol, UK. [4]University Hospital Southampton, Southampton, UK. [5]Faculty of Medicine, University of Southampton, Southampton, UK. [6]Department of Medical Psychology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [7]Institute of Psychology, Leiden University, Leiden, Netherlands

Contact address: Tony Kendrick, Primary Care and Population Sciences, Faculty of Medicine, Aldermoor Health Centre, University of Southampton, Aldermoor Close, Southampton, SO16 5ST, UK. ark1@soton.ac.uk.

Editorial group: Cochrane Common Mental Disorders Group.
Publication status and date: Edited (no change to conclusions), comment added to review, published in Issue 8, 2016.
Review content assessed as up-to-date: 18 May 2015.

Citation: 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 Database of Systematic Reviews* 2016, Issue 7. Art. No.: CD011119. DOI: 10.1002/14651858.CD011119.pub2.

ABSTRACT

Background

Routine outcome monitoring of common mental health disorders (CMHDs), using patient reported outcome measures (PROMs), has been promoted across primary care, psychological therapy and multidisciplinary mental health care settings, but is likely to be costly, given the high prevalence of CMHDs. There has been no systematic review of the use of PROMs in routine outcome monitoring of CMHDs across these three settings.

Objectives

To assess the effects of routine measurement and feedback of the results of PROMs during the management of CMHDs in 1) improving the outcome of CMHDs; and 2) in changing the management of CMHDs.

Search methods

We searched the Cochrane Depression Anxiety and Neurosis group specialised controlled trials register (CCDANCTR-Studies and CCDANCTR-References), the Oxford University PROMS Bibliography (2002-5), Ovid PsycINFO, Web of Science, *The Cochrane Library*, and International trial registries, initially to 30 May 2014, and updated to 18 May 2015.

Selection criteria

We selected cluster and individually randomised controlled trials (RCTs) including participants with CMHDs aged 18 years and over, in which the results of PROMs were fed back to treating clinicians, or both clinicians and patients. We excluded RCTs in child and adolescent treatment settings, and those in which more than 10% of participants had diagnoses of eating disorders, psychoses, substance use disorders, learning disorders or dementia.

Data collection and analysis

At least two authors independently identified eligible trials, assessed trial quality, and extracted data. We conducted meta-analysis across studies, pooling outcome measures which were sufficiently similar to each other to justify pooling.

Main results

We included 17 studies involving 8787 participants: nine in multidisciplinary mental health care, six in psychological therapy settings, and two in primary care. Pooling of outcome data to provide a summary estimate of effect across studies was possible only for those studies using the compound Outcome Questionnaire (OQ-45) or Outcome Rating System (ORS) PROMs, which were all conducted in multidisciplinary mental health care or psychological therapy settings, because both primary care studies identified used single symptom outcome measures, which were not directly comparable to the OQ-45 or ORS.

Meta-analysis of 12 studies including 3696 participants using these PROMs found no evidence of a difference in outcome in terms of symptoms, between feedback and no-feedback groups (standardised mean difference (SMD) -0.07, 95% confidence interval (CI) -0.16 to 0.01; P value = 0.10). The evidence for this comparison was graded as low quality however, as all included studies were considered at high risk of bias, in most cases due to inadequate blinding of assessors and significant attrition at follow-up.

Quality of life was reported in only two studies, social functioning in one, and costs in none. Information on adverse events (thoughts of self-harm or suicide) was collected in one study, but differences between arms were not reported.

It was not possible to pool data on changes in drug treatment or referrals as only two studies reported these. Meta-analysis of seven studies including 2608 participants found no evidence of a difference in management of CMHDs between feedback and no-feedback groups, in terms of the number of treatment sessions received (mean difference (MD) -0.02 sessions, 95% CI -0.42 to 0.39; P value = 0.93). However, the evidence for this comparison was also graded as low quality.

Authors' conclusions

We found insufficient evidence to support the use of routine outcome monitoring using PROMs in the treatment of CMHDs, in terms of improving patient outcomes or in improving management. The findings are subject to considerable uncertainty however, due to the high risk of bias in the large majority of trials meeting the inclusion criteria, which means further research is very likely to have an important impact on the estimate of effect and is likely to change the estimate. More research of better quality is therefore required, particularly in primary care where most CMHDs are treated.

Future research should address issues of blinding of assessors and attrition, and measure a range of relevant symptom outcomes, as well as possible harmful effects of monitoring, health-related quality of life, social functioning, and costs. Studies should include people treated with drugs as well as psychological therapies, and should follow them up for longer than six months.

PLAIN LANGUAGE SUMMARY

Using patient-reported outcome measures to monitor progress among adults with common mental health disorders

Why is this review important?

One in six people suffer from a common mental health disorder (CMHD), including depression and anxiety disorders. Patient reported outcome measures (PROMs) are questionnaires on patients' symptoms, functioning, and relationships. Using PROMs to monitor the progress of people with CMHDs might improve treatment outcomes, and change the management of CMHDs.

Who will be interested in this review?

People with CMHDs; health professionals in primary care, psychological therapy and mental health services; health service commissioners.

What questions does this review aim to answer?

Does the use of PROMs to monitor progress in people with CMHDs improve health outcomes, including symptoms, quality of life, and social functioning?

Does the use of PROMs in people with CMHDs change the way their problems are managed, including drug therapy and referrals for specialist help?

Which studies were included in the review?

Trial databases were searched to find all high-quality studies of the use of PROMs to monitor the treatment of CMHDs published up to May 2015. Included studies had to be randomised controlled trials in adult participants, where the majority diagnosed had a CMHD.

Seventeen studies involving 8787 participants were included in the review, nine from mental health, six from psychological therapy, and two from primary care settings.

The quality of the studies was rated 'low' to 'moderate'.

What does the evidence from the review tell us?

Routine outcome monitoring of CMHDs using PROMs was not shown conclusively to be helpful in analyses combining study results, either in terms of improving patient symptom outcomes (across 12 studies), or in changing the duration of treatment for their conditions (across seven studies). It was not possible to analyse changes in drug treatment or referrals for further treatment as only two studies reported these. Similarly, health-related quality of life, social functioning, adverse events, and costs were reported in very few studies.

What should happen next?

More research of better quality is required, especially in primary care where most CMHDs are treated. Studies should include people treated with drugs as well as psychological therapies, and should follow them for longer than six months. As well as symptoms and length of treatment, studies should measure possible harms, quality of life, social functioning, and the costs of monitoring.

SUMMARY OF FINDINGS FOR THE MAIN COMPARISON [Explanation]

Feedback of PROM scores for routine monitoring of common mental health disorders

Patient or population: People with common mental health disorders[1]
Settings: Primary care, multidisciplinary mental health care, or psychological therapies
Intervention: Feedback of PROM scores to clinician, or both clinician and patient
Comparator: No feedback of PROM scores

Outcomes and length of follow-up	Illustrative risk		Number of participants (number of studies)	Quality of the evidence (GRADE)	Comments
	Assumed risk (range of means in no-feedback groups)	**Relative effect** (95% CI) in feedback groups			
Mean improvement in symptom scores Outcome Questionnaire-45 (OQ-45) or Outcome Rating Scale (ORS) Follow-up: 1-6 months[2]	Mean scores in no-feedback groups ranged from **51.8 to 101.5 points for OQ-45** and from **23.8 to 29.5 points for ORS.** Standard deviations ranged from 17.8 to 28.6 points for OQ-45 and from 7.1 to 9.6 points for ORS	Standard mean difference in symptom scores at end of study in feedback groups was **0.07 standard deviations lower** (0.16 lower to 0.01 higher) [3,4]	3696 (12 studies)	⊕⊕◯◯ low[5,6]	Neither study in the primary care setting used the OQ-45 or ORS PROMs, and so could not be included in this meta-analysis
Health-related quality of life Medical Outcomes Study Short Form (SF-36) Follow-up: 1-5 months[2] Medical Outcomes Study (SF-12) physical and mental subscales) Scale from 0-100 Follow-up: 0-1 year	Study results could not be combined in a meta-analysis as data were not available in an appropriate format Mathias 1994 reported no significant differences between feedback and control groups on all nine domains of the SF-36 Scheidt 2012 reported no significant differences between feedback and no-feedback groups in physical or mental sub-scale scores		583 (1 study) 587 (1 study)	⊕⊕⊕◯ moderate[7]	
Adverse events PHQ-9 questionnaire[8] Follow-up: 6 months	Chang 2012 reported no immediate suicide risk across both feedback and no-feedback groups combined. Number per group not given		642 (1 study)	⊕⊕⊕◯ moderate[7]	

Social functioning Follow-up: 0-1 year[2]	Data for the social functioning subscale of the OQ-45 were considered separately in Hansson 2013 and no difference was found		262 (1 study)	⊕⊕◯◯ low[9]	
Costs	Not estimable		0 (0 studies)		No study assessed the impact of the intervention on direct or indirect costs
Changes in the management of CMHDs Changes in drug therapy and referrals for specialist care Follow-up: 1-6 months[2]	Study results could not be combined in a meta-analysis as data were not available in an appropriate format Chang 2012 and Mathias 1994 both reported no significant differences in changes in drug therapy between study arms Mathias 1994 reported mental health referrals were significantly more likely in the feedback group (OR 1.73, 95% CI 1.11 to 2.70)		1215 (2 studies)	⊕⊕⊕◯ moderate[7]	
Changes in the management of CMHDs Number of treatment sessions received Follow-up: 1-6 months[2]	Mean in no-feedback groups ranged from **3.7 to 33.5 treatment sessions**	Mean difference in number of treatment sessions in feedback groups was **0.02 lower** (0.42 lower to 0.39 higher)	2608 (7 studies)	⊕⊕◯◯ low[10]	Post-hoc analysis. Changes in medication and referrals for additional therapy were not assessed by any of these studies

CI: Confidence interval

GRADE Working Group grades of evidence
High quality: Further research is very unlikely to change our confidence in the estimate of effect.
Moderate quality: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
Low quality: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
Very low quality: We are very uncertain about the estimate.

[1]Studies were included if the majority of people diagnosed had CMHDs and no more than 10% had diagnoses of psychotic disorders, learning difficulties, dementia, substance misuse, or eating disorders
[2]Duration of therapy was variable in all studies and determined by the clinician or the patient, or both
[3]OQ-45 range of scores 0-180 (0 best, 180 worst). Three studies (Murphy 2012, Reese 2009a and Reese 2009b) used the ORS - range of scores 0-40 (0 worst, 40 best)
[4] This is a difference in standard deviations. A standard deviation of 0-0.2 represents no to small difference between groups (rule of thumb according to Cohen's interpretation of effect size)

[5]An expected effect size of 0.3 SD would require a minimum total sample size of 352 participants. An expected effect size of 0.1 SD would require 3142 participants

[6]Downgraded two levels due to risk of bias (all included studies were judged at high risk of bias in at least two domains, in particular blinding of participants and outcome assessment, and attrition), and indirectness (although symptom scores were compared between feedback and non-feedback groups, wider social functioning and quality-of-life measurements were not assessed in nearly all studies)

[7]Downgraded one level due to risk of bias (judged at high risk of bias in at least two domains, in particular blinding of participants and outcome assessment, and attrition)

[8]Number of PHQ-9 questionnaires which contained reports of self-harming thoughts

[9]Downgraded two levels due to risk of bias and imprecision, as total participant numbers were less than 400

[10]Downgraded two levels due to risk of bias and for imprecision: estimate of effect includes no effect and incurs very wide confidence intervals

BACKGROUND

Description of the condition

Common mental health disorders (CMHDs) are prevalent, often very disabling and very costly. They include depression (including major depression, dysthymia and minor or mild depression); mixed anxiety and depression; and specific anxiety disorders, namely generalised anxiety disorder (GAD), phobias, obsessive-compulsive disorder (OCD), panic disorder and post-traumatic stress disorder (PTSD) (McManus 2009). Katon and Schulberg estimated in 1992 that depression fulfilling the criteria for major depression in the American Psychiatric Association Diagnostic and Statistical Manual, 4th edition (DSM-IV) (APA 2000) occurred in 2% to 4% of people in the community, 5% to 10% of primary care patients, and 10% to 14% of medical inpatients; but in each setting there were two to three times as many people with depressive symptoms that were short of the major depression criteria (Katon 1992). Prevalence rates of major depression of 13.9% in women and 8.5% in men, and of anxiety disorders of 10% and 5% respectively, have been found in family practice attendees across Europe (King 2008). The estimated one-week prevalence of CMHDs among adults in England in 2007, according to the criteria of the World Health Organization's International Classification of Diseases (ICD-10) (WHO 1992) was found to be 17.6%, including mixed anxiety and depression in 9.7%; GAD in 4.7%; depressive episode in 2.6%; phobia in 2.6%; OCD in 1.3%; and panic disorder in 1.2% (McManus 2009). In the US National Comorbidity Survey, lifetime prevalence estimates were 16.6% for DSM-IV major depression; 6.8% for PTSD; 5.7% for GAD; 4.7% for panic disorder; 2.5% for dysthymia; 1.6% for OCD; and 1.4% for agoraphobia (Kessler 2005).

Depression is often chronic and relapsing, resulting in high levels of disability and poor quality of life (Wells 1989), generally high levels of health service use and associated economic costs (Simon 1997), and death from suicide in between 2% and 8% of cases (Bostwick 2000). Major depressive disorder appears to be increasing in prevalence (Compton 2006) and in the Global Burden of Disease Study 2010 (Murray 2010) has moved up to 11th from 15th in the ranking of disorders according to burden in terms of disability adjusted life years (a 37% increase), becoming the second leading cause of years lived with disability, due to population growth and ageing (Ferrari 2013).

The King's Fund estimated that in the UK 1.45 million people would have depression by 2026, and the total cost to the nation would exceed GBP 12 billion per year, including prescriptions, inpatient and outpatient care, supported accommodation, social services and lost employment (McCrone 2008). The total medical and productivity costs per person with any anxiety disorder were estimated to be around USD 6500 in the USA in 1999 (Marciniak 2004), and across Europe the annual costs of anxiety disorders, including health service costs, welfare benefits and lost productivity, were estimated to exceed USD 40 billion in 2004 (Andlin-Sobocki 2005).

Depression is usually treated in primary care with selective serotonin reuptake inhibitor (SSRI) antidepressant drugs (in around 80% of cases), psychological treatments (in around 20%), or both (Kendrick 2009); and in one-third to one-half of people with major depression, the symptoms persist over a six to 12-month period (Gilchrist 2007; Katon 1992). Evidence-based guidelines recommend psychological treatments such as cognitive-behaviour therapy (CBT) as first-line treatment for anxiety disorders (NICE 2011a) but SSRIs are also frequently prescribed for their treatment, often because psychological treatments are not available. It is recommended that people prescribed antidepressants are seen for regular follow-up during treatment. For example, the UK National Institute for Health and Care Excellence (NICE) 2009 guideline on the management of depression in adults recommended that people started on antidepressants who were not considered to be at increased risk of suicide should normally be seen after two weeks, then at intervals of two to four weeks in the first three months, and then at longer intervals if their response to treatment was good (NICE 2009). At each visit clinicians were recommended to evaluate response (symptoms and functioning), adherence to treatment, drug side-effects and suicide risk (NICE 2009). This evaluation is usually based on clinical judgement alone, but in recent years clinicians have been advised to consider using patient reported outcome measures (PROMs) to augment their clinical judgement. NICE guidance states all staff carrying out the assessment of common mental health disorders should be competent in the use of formal assessment measures and routine outcome measures (NICE 2011a).

Description of the intervention

PROMs assess patients' experiences of their symptoms, their functional status and their health-related quality of life. So they can help to determine the outcome of care in terms of these aspects from the patient's perspective as an expert in the lived experience of their own health. PROMs are different to measures of patients' experience of, or satisfaction with, the care they receive (Black 2013). PROMs are often self-report measures that should therefore be free of observer rating bias, but they can also be interview-based measures that involve the interviewer in interpreting the patients' responses to questions.

The treatment of CMHDs has been augmented in a number of studies by administering PROMs measuring symptoms of depression or anxiety, social functioning or health-related quality of life, and feeding the results back to the treating clinician or both the treating clinician and the patient. Feedback of the results is the essential element. The intervention will usually include education of the clinician, or both the clinician and patient, about the measures used and their interpretation. It may or may not also include

specific instructions on action to take in light of the results, which may be in the form of an algorithm.

How the intervention might work

Carlier 2012 identifies two main theories concerning the links between the use of PROMs, the process of care and outcomes for patients, Feedback Intervention Theory (FIT) and Therapeutic Assessment (TA). FIT suggests that feedback of the results of PROMs to healthcare professionals influences them to adjust treatment or refer for alternative interventions, improving care when measured against best practice guidelines; while TA focuses on the potential therapeutic effects of feeding back the test results to patients.

Greenhalgh 2005 pointed out that feedback to the clinician may initiate specific changes in management, including ordering further tests, referring to other professionals, changing treatments, and giving advice and education to the patient on better control or management of the problem. Feeding the results back to the patient as well as to the clinician can potentially further improve the process of care, as patients often like to be more involved in their own care, which may be beneficial in itself. This may promote better communication and a greater understanding of the patient's personal circumstances, enabling joint decision-making between clinician and patient, increasing concordance and patient adherence to treatment through agreeing shared goals, and increasing patient satisfaction, all of which in turn can potentially improve the outcome for the patient.

Observational studies suggest that general practitioner (GP) treatment decisions (to prescribe antidepressants, to subsequently change prescriptions, or refer patients for specialist treatment) might be influenced by the results of patient-completed depression symptom questionnaires at diagnosis (Kendrick 2009) and follow-up (Moore 2012), in line with the predictions of FIT. A trial of feeding back depression symptom questionnaire scores to primary care physicians and patients in the USA led to increased rates of response to treatment and remission among patients in the intervention arm (Yeung 2012) although this was despite an apparent lack of significant changes in the physicians' management of the patients' depression (Chang 2012). The authors suggested that frequent symptom measurement might have increased patients' symptom awareness and their ability to report relevant symptoms to their physicians, or made them feel more supported, contributing to a lower medication discontinuation rate in the intervention group. Qualitative research suggests that patients with depression do value the use of symptom questionnaires to assess their condition (Dowrick 2009) and the effectiveness of their treatment (Malpass 2010). It might be that if patients feel that they have been assessed more thoroughly and become more involved in the care of their disorder through the completion of PROMs, together with feedback of the significance of the results, this can help them to improve more quickly even in the absence of significant changes in management, in line with the predictions of TA.

Why it is important to do this review

The use of PROMs has been promoted in recent years as a way for patients to become more involved in their own care and to help health professionals make better decisions about their treatments (Black 2013; Black 2015; Fitzpatrick 2009).

In particular, the use of PROMs in depression has been promoted in important policy pronouncements. The US Federal Health Resources and Services Administration (HRSA) Collaborative on Depression included quality standards for the proportion of patients assessed using the self-complete Patient Health Questionnaire (PHQ-9) depression symptom measure (Spitzer 1999) at diagnosis and follow-up (HRSA 2005). The NICE 2009 depression guideline recommended that clinicians should consider using a validated measure (for example for symptoms, functions and disability) to inform and evaluate treatment (NICE 2009). The subsequent NICE quality standard on assessment of depression recommended that practitioners delivering interventions for people with depression should record the results of validated health outcome measures at each treatment contact and use the findings to adjust their delivery of interventions (NICE 2011b). In 2009 a performance indicator was added to the UK National Health Service (NHS) GP pay for performance scheme (the Quality and Outcomes Framework or QOF), financially incentivising the follow-up assessment of depression with symptom questionnaires five to 12 weeks after diagnosis (BMA & NHS Employers 2009). The UK NHS Increasing Access to Psychological Therapies (IAPT) programme, extending the provision of psychological treatments for CMHDs nationwide, adopted an information standard with an instruction to record PROMs at every visit, including the PHQ-9 for depression, the self-complete Generalised Anxiety Disorder questionnaire (GAD-7) for anxiety (Spitzer 2006), and the Work and Social Adjustment Scale (WSAS, Mundt 2002) for social functioning (IAPT 2011).

The potential for PROMs to improve the care and self-care of CMHDs cannot be assumed however. The administration of symptom, social functioning, or quality-of-life questionnaires to each and every patient with a CMHD adds up to a significant investment of resources in terms of professionals' time given the high numbers of patients with CMHDs, especially in primary care. Following the introduction of the QOF performance indicator financially incentivising the follow-up assessment of depression with symptom questionnaires, GPs in the UK reported completing more than 1.1 million follow-up assessments between April 2009 and March 2013 (74% of 1.5 million eligible cases identified in those five years) (QOF Database 2013). The cost to the NHS of those assessments added up to more than GBP 25 million per year in terms of GP time and the incentive payments. Therefore, even such relatively simple quality improvement strategies should be supported by evidence of clinical benefit and cost-effectiveness.

There have been a number of previous systematic reviews related to this question including studies in different sectors of health care: one of studies in non-psychiatric settings (Gilbody 2002); one of

studies in clinical psychology practice (Lambert 2003), updated in 2010 (Shimokawa 2010); two combining studies in multidisciplinary mental health care (which we previously referred to as 'specialist psychiatric practice', see section on Differences between protocol and review) and clinical psychology practice (Davidson 2014; Knaup 2009); and one limited to studies in primary care (Shaw 2013). The review by Gilbody and colleagues failed to show an impact of patient-centred outcome instruments assessing patient needs or measuring quality of life in non-psychiatric settings (Gilbody 2002). However, Knaup and colleagues' systematic review of studies in specialist psychological and multidisciplinary mental health care settings, which included the studies previously reviewed by Lambert and colleagues (Lambert 2003), was more positive, demonstrating benefits of routine outcome measurement for a range of mental health problems (Knaup 2009). Outcomes were found to be improved with an effect size of between 0.1 and 0.3 standard deviations, being improved more when patients were involved in rating their own problems and received feedback on their progress in addition to feedback to the practitioner (Knaup 2009). However, this review included studies of people with more severe mental illnesses as well as CMHDs. Conversely, the 2013 review (Shaw 2013) had a narrow focus as it was limited to studies of the assessment and monitoring of depression in primary care using questionnaires recommended in the NHS GP contract QOF, namely the PHQ-9, Hospital Anxiety and Depression Scale (HADS) (Zigmond 1983) and Beck Depression Inventory (BDI) (Beck 1961) or BDI-II (Beck 1996). Other systematic reviews and meta-analyses have included studies of the use of PROMs as screening or diagnostic tools together with studies of their use as follow-up monitoring measures (Carlier 2012; Poston 2010) or have included studies of the use of PROMs in the management of physical disorders together with studies in mental health care (Boyce 2013; Marshall 2006; Valdera 2008). One recent systematic review included only studies which evaluated feeding back the results of PROMs in terms of changes in the particular PROM score rather than other relevant outcome measures (Boyce 2013). There has been no systematic review of the use of PROMs in the routine outcome monitoring of CMHDs in adults across primary care, psychological therapy, and multidisciplinary mental health care settings. Given the high prevalence of CMHDs, the current policy drive promoting routine outcome monitoring across these settings, and the likely significant cost of such widespread monitoring of highly prevalent conditions, there is an urgent need for evidence to guide further developments in policy and clinical practice. We therefore aimed to conduct a comprehensive, up-to-date systematic review of the use of PROMs in CMHDs, including studies across primary care, multidisciplinary mental health care, and psychological therapy settings. We aimed to include measures of social functioning and health-related quality of life (QoL) as well as measures of symptoms of depression and anxiety, because functioning and QoL measures may also influence clinician treatment decisions or patient involvement in their own care, or both, and therefore outcomes for patients.

PROMs can be used as a tool to identify patients with CMHDs whose problems would otherwise be missed, but in this review we were not concerned with the use of PROMs as a screening tool. This was the subject of a previous review, Gilbody 2008. In this review we were concerned with the use of PROMS in monitoring patients' progress and response to treatment, which requires feedback and assessment of the results at follow-up, after a period of treatment, rather than screening or assessment only before diagnosis or at the point of diagnosis.

We conducted this review according to the methods set out in the protocol (Kendrick 2014).

OBJECTIVES

To assess the effects of routine measurement and feedback of the results of PROMs during the management of CMHDs in 1) improving the outcome of CMHDs; and 2) in changing the management of CMHDs.

METHODS

Criteria for considering studies for this review

Types of studies

We included randomised controlled trials (RCTs), including cluster RCTs and RCTs randomised at the level of individual participants. We excluded non-randomised trials.

We planned to include cluster trials where clusters were allocated to intervention or control arms using a quasi-randomised method, such as minimisation, to avoid significant imbalance between arms arising by chance when the number of clusters is relatively small, but planned to exclude quasi-randomised trials where allocation was at the level of individual participants. We planned to exclude cross-over trials because of the very high risk of carry-over of the intervention into the control arm after participating clinicians or patients cross over. We also planned to exclude uncontrolled before and after trials, and observational studies. However, none of these types of studies was identified.

Types of participants

Participant characteristics

We selected studies which included participants with common mental health disorders (CMHDs) aged 18 years and over, of both genders and all ethnic groups. We excluded studies in child and

adolescent treatment settings, as the diagnostic categories included within the group recognised as CMHDs are limited to adults, and in addition the presence of a parent or other carer accompanying a child or adolescent patient complicates the issues of who is providing responses to PROMs administered to monitor the outcome of treatment, and to whom feedback of the results is given.

Diagnosis

We included adult patients with any CMHD, including both those with formal diagnoses according to the criteria of the DSM (APA 2000) or ICD (WHO 1992), and those diagnosed through clinical assessment only, unaided by formal reference to specific diagnostic criteria. The specific disorders included were:

1. depression (including major depression, dysthymia, and minor or mild depression);
2. mixed anxiety and depression;
3. generalised anxiety disorder (GAD);
4. phobias;
5. obsessive-compulsive disorder (OCD);
6. panic disorder;
7. post-traumatic stress disorder (PTSD);
8. adjustment reaction.

We included studies in which the diagnoses of the majority of participants were reported as CMHDs, even if a proportion of participants were not given a specific diagnosis, or were reported as having relationship or interpersonal difficulties, 'somatoform disorders', 'other' diagnoses not further specified, or 'administrative codes'. This was a change from the protocol as we planned originally to include only studies with participants specifically diagnosed with one of the disorders listed above, but after discussion within the review study group we decided to include these studies in order to be able to include studies which had a majority of participants diagnosed with CMHDs (see section on Differences between protocol and review).

We excluded studies with more than 10% of patients diagnosed specifically with psychoses, substance use disorders, learning disorders or dementia. We also excluded studies with more than 10% of participants diagnosed with eating disorders, as they are a separate group of disorders not usually included within the group recognised as CMHDs, and the PROMs used for eating disorders are less generic and specifically concentrate on eating habits and weight control measures. This was also a change from the protocol as we planned to exclude studies with any participants at all in these categories, but again, after discussion within the review study group we decided to include studies with fewer than 10% of participants with these diagnoses, in order once again to be able to include studies which had a majority of participants with CMHDs (see section on Differences between protocol and review).

Where studies did not report the diagnoses of participants, we attempted to contact the authors to request information on the participant diagnoses, and whether they would have met the review inclusion and exclusion criteria. This was an addition to the protocol (see section on Differences between protocol and review).

We carried out sensitivity analyses omitting studies which did not report specific diagnoses of CMHDs for 20% or more of their participants, to determine whether these decisions affected the findings.This was an addition to the protocol agreed once again after discussion within the review study group (see section on Differences between protocol and review).

Co-morbidities

Participants diagnosed with or without co-morbid physical illnesses were included to ensure as representative a sample as possible.

Setting

Three settings were included: primary care (where the clinicians were all primary care physicians and available treatments post-assessment included either drug therapy or referral for psychological therapy); multidisciplinary mental health care (where the clinicians included psychiatrists, psychologists, mental health social workers or mental health nurses, and available treatments included drugs, psychological therapies, and physical treatments); and psychological therapies (where the clinicians were psychologists, social workers or nurses and available treatments were all psychological).

Subset data

We planned to include trials that provided data on a relevant subset of their participants, for example studies which compared usual care in one arm with routine outcome monitoring in another, even if there was a third arm with a more complex intervention, but we did not identify any such trials. We also planned to include trials that included a subset of participants who met our criteria for the review, for example in terms of the types of disorder or age range, if the data for those participants could be extracted separately from the rest of the trial sample, but again we did not identify any such trials.

Types of interventions

Experimental intervention

The intervention consisted of augmenting the assessment and management of CMHDs by both of the following.

1. Measuring patient reported outcomes (PROMs), including self-complete or administered measures of:

i) depressive symptoms, for example the PHQ-9 (Spitzer 1999). We planned to include the HADS depression subscale (HAD-D) (Zigmond 1983); BDI (Beck 1961) and BDI-II (Beck

1996), but found no relevant studies which used them as PROMs;

ii) anxiety symptoms, for example the Beck Anxiety Inventory (BAI) (Wetherall 2005). We planned to include the GAD-7 (Spitzer 2006) but no trials used it;

iii) health-related QoL, for example with the Medical Outcomes Study Short Form SF-36 (Wells 1989) or SF-12 (Ware 1996). We planned to include the EuroQol five item EQ-5D questionnaire (Dolan 1997) but no trials used it;

iv) symptoms, individual functioning, and social functioning as composite measures, for example the 45-item Outcomes Questionnaire (OQ-45) (Lambert 2004), and the Outcome Rating Scale (ORS) (Miller 2003). We planned to include the Clinical Outcomes in Routine Evaluation Outcome Measure (CORE-OM) (Barkham 2006) but no trials used it

2. Feeding the results back to the treating clinician, to both the clinician and the patient, or to the patient only.

We also planned to include studies using the following as PROMs but found no relevant studies:

1. measures of depression and anxiety combined, for example the self-complete General Health Questionnaire (GHQ-28) (Goldberg 1972) or the administered Mini-International Neuropsychiatric Interview (MINI) (Sheehan 1998); and
2. measures of social functioning, for example the WSAS (Mundt 2002) or the Social Adjustment Scale (SAS) (Cooper 1982)).

Comparator intervention

The comparator was usual care for CMHDs without feeding back the results of PROMs. Routine care includes usual patient-clinician interaction with non-standardised history-taking, investigation, referral, intervention and follow-up. Trials were excluded if the comparator interventions involved the use of feedback of the results of PROMs as a clinical tool to inform management of the participants. Measures of depression, anxiety, social functioning and quality of life may have been assessed independently by researchers in both the intervention and control conditions to determine the effects of the intervention, but the active component, which was the feeding back of this information to the clinician, or to the patient, or to both clinician and patient, had to occur only in the intervention arm.

Excluded interventions

We excluded studies where the intervention arm was subject to additional components over and above the feedback of PROM results, including pharmacological or psychological treatments that were not available to both the intervention and control groups. A number of more complex interventions have been advocated to improve the quality of care of people with CMHDs including case management (Simon 2004) and collaborative care (Archer 2012), and these usually include feeding back the results of PROMs at initial assessment and follow-up to inform treatment. However, this review was limited to the effects of feedback of the results of PROMs alone, rather than their use as a component of complex interventions which also enhanced the process of care through case management, collaborative care, active outreach or other systems or processes over and above usual care. It would not have been possible to distinguish the effects of outcome monitoring from other active components in such studies.

Types of outcome measures

Studies that met the above inclusion criteria were included regardless of whether they reported on the following outcomes.

Primary outcomes

1. Mean improvement in symptom scores

Mean improvement in symptom scores (and standardised effect size) from baseline to follow-up on a symptom-specific scale, which was either:

1. an interviewer-rated measure; or
2. a self-complete questionnaire measure.

Measures used included:

1. interviewer-rated measures of depression and anxiety including the Diagnostic Interview Schedule (DIS) for DSM-III disorders (Robins 1981); and
2. self-complete measures including the PHQ-9 (Spitzer 1999); BDI (Beck 1961) and BDI-II (Beck 1996) for depression; the BAI (Wetherall 2005) for anxiety; and the Hopkins symptom checklist SCL-90 (Derogatis 1974; Derogatis 1983) for both anxiety and depression.

We also planned to include, but found no relevant studies which used the following as primary outcome measures:

1. the interviewer-rated Hamilton Depression Rating Scale (HDRS or HAMD) (Hamilton 1960); Montgomery-Asberg Depression Rating Scale (MADRS) (Montgomery 1979); Structured Clinical Interview for DSM-IV disorders (SCID) (First 1997); and the interviewer-rated version of the Quick Inventory of Depressive Symptomatology (QIDS) (Trivedi 2004);
2. the self-complete Community Epidemiologic Survey Depression (CES-D) scale for DSM-III depression (Radloff 1997); Zung depression scale (SDI) (Zung 1965); GAD-7 anxiety scale (Spitzer 2006); GHQ (Goldberg 1972); HADS (Zigmond 1983); Hopkins symptom checklist (Derogatis 1974; Derogatis 1983); Clinical Interview Schedule, Revised (CIS-R) for ICD-10 disorders (Lewis 1992); and the self-complete version of the Quick Inventory of Depressive Symptomatology (QIDS) (Trivedi 2004).

2. Health-related quality of life

Health-related quality of life, assessed using specific measures at baseline and follow-up, including the SF-36 (Wells 1989). We also planned to include the EQ-5D (Dolan 1997) but identified no relevant trials which used it.

3. Adverse events, including:

1. numbers and types of antidepressant drug side-effects;
2. numbers of incidences of self-harm, and
3. numbers of suicides.

Secondary outcomes

4. Changes in the management of CMHDs

Changes in the management of CMHDs following administration and feedback of the results of PROMs, including:

1. number of changes in drug prescribing (a new prescription, a change in dose or type of drug, or the ending of a prescription);
2. number of referrals for psychological assessment or treatment;
3. number of referrals for psychiatric assessment or treatment.

These are relevant secondary outcomes, as they indicate more proactive care, which might lead to more positive outcomes, although a change in management cannot by itself be regarded as necessarily a positive outcome.

5. Social functioning

Social functioning assessed using specific measures at baseline and follow-up, for example the WSAS (Mundt 2002). We also planned to include the SAS (Cooper 1982) but identified no relevant trials which used it.

6. Costs, including:

1. the direct costs of administering PROMs and delivering feedback of the results;
2. costs to the health service, including consultations, prescriptions, outpatient attendances and hospital admissions; and
3. societal costs, including costs to the patient and to society in terms of loss of employment and costs of sickness benefits.

Timing of outcome assessment

We planned to divide the reporting of research outcomes into:

1. short-term, up to six months after baseline assessment; and
2. long-term, beyond six months.

Hierarchy of outcome measures

We planned to select self-complete research outcome measures in preference to interviewer-rated measures of symptoms, social functioning or health-related quality of life as they are less prone to detection bias due to unblinding of the researcher assessing the outcome. In completing an interviewer-rated measure the researcher filters all patient reported responses, while for self-complete measures only those responses which the patient chooses to discuss with the researcher can be influenced by an unblinded researcher.

Search methods for identification of studies

The Cochrane Depression, Anxiety and Neurosis Review Group's Specialised Register (CCDANCTR)

The Cochrane Depression, Anxiety and Neurosis Group (CCDAN) maintain two clinical trials registers at their editorial base in Bristol, UK: a references register and a studies-based register. The CCDANCTR-References Register contains over 39,000 reports of RCTs in depression, anxiety and neurosis. Approximately 50% of these references have been tagged to individual, coded trials. The coded trials are held in the CCDANCTR-Studies Register and records are linked between the two registers through the use of unique Study ID tags. Coding of trials is based on the EU-Psi coding manual using a controlled vocabulary (please contact the CCDAN Trials Search Co-ordinator for further details). Reports of trials for inclusion in the Group's registers are collated from routine (weekly), generic searches of Ovid MEDLINE (1950 -), EMBASE (1974 -) and PsycINFO (1967 -); quarterly searches of the Cochrane Central Register of Controlled Trials (CENTRAL); and review specific searches of additional databases. Reports of trials are also sourced from international trials registers through the World Health Organization's trials portal (the Clinical Trials Registry Platform (ICTRP)) and the handsearching of key journals, conference proceedings and other (non-Cochrane) systematic reviews and meta-analyses.

Details of CCDAN's generic search strategies (used to identify RCTs) can be found on the Group's website.

Electronic searches

1. The CCDANCTR (References and Studies Register) was initially searched to 30 May 2014 using the following terms:
#1 ("affective disorder*" or "common mental disorder*" or "mental health" or "acute stress" or adjustment or anxi* or compulsi* or obsess* or OCD or depressi* or dysthymi* or neurosis or neuroses or neurotic or panic or *phobi* or PTSD or posttrauma* or "post trauma*" or "stress disorder*" or trauma* or psychotrauma*): ti,ab,kw,ky,emt,mh,mc
#2 PROMS

#3 ("patient reported outcome*" or "patient reported assessment*" or "patient reported symptom*")
#4 "patient outcome*"
#5 ((patient* or client* or tailored) NEAR2 feedback)
#6 (patient* NEXT ("self assess*" or "self report" or "self monitor*"))
#7 (patient* NEAR2 progress*)
#8 "client report*"
#9 ((active or routine* or regular*) NEAR2 (feedback or measurement* or monitor*))
#10 (monitor* and feedback*)
#11 ("feedback to" or "feed back to" or "fed back to"):ab
#12 ((symptom* or treatment) NEXT monitor*)
#13 (monitor* NEAR2 ("common mental disorder*" or anxi* or compulsi* or obsess* or OCD or depressi* or neurosis or neuroses or neurotic or panic or *phobi* or PTSD or posttrauma* or "post trauma*" or "acute stress" or "stress disorder*" or trauma*))
#14 ((follow-up* or "follow up*") and assess*):ti
#15 (needs NEAR3 assess*)
#16 (outcome* NEAR (clinical or feedback or manag* or monitor*)):ti
#17 "severity questionnaire*"
#18 severity:ti,kw,ky and (assess* or measure* or outcome* or questionnaire* or score*):ti
#19 ("case management" or "enhanced care")
#20 (#2 or #3 or #4 or #5 or #6 or #7 or #8 or #9 or #10 or #11 or #12 or #13 or #14 or #15 or #16 or #17 or #18 or #19)
#21 (#1 and #20)
[Key. ab:abstract; emt:EMTREE headings; kw:CRG keywords; ky: other keywords; mc:MesH check words; mh:MeSH headings]

Due to the nature of the intervention (patient reported outcome measures) the search strategy was designed to favour specificity (precision) over sensitivity (recall of all potentially relevant reports). A sensitive search would retrieve too much noise as most of the measures and questionnaires under review are much more frequently used to assess symptom severity or quality of life as research outcomes in treatment trials in patients with CMHDs than as PROMs used for clinical assessment.

2. Complementary searches were conducted on the following bibliographic databases using relevant subject headings (controlled vocabularies) and search syntax that were appropriate to each resource. Searches initially performed to 5 June 2014:

(i) Ovid PsycINFO (all years)

Although PsycINFO is routinely searched to inform the CCDANCTR, we conducted an additional search of this database to increase the sensitivity of our search methods, adding wait-list control and treatment-/care-as-usual to CCDAN's standard RCT filter. The search strategy is described in Appendix 1.

(ii) PROM Bibliography database (all years to 2005)

The PROM Bibliography was searched for RCTs in mental health. This database, which is available through The Patient-Reported Outcomes Measurement Group at the University of Oxford, was first published in 2002 with funding from the Department of Health (DH). It was further developed with DH funding to 2005 and contains over 16,000 records relating to patient reported outcome measures.

(iii) Web of Science (WoS): Science Citation Index (cited reference search, all years as appropriate)

3. International trial registries were also searched on 19 February 2015 and 9 April 2015 via the World Health Organization's trials portal (ICTRP) and ClinicalTrials.gov to identify unpublished or ongoing studies. We searched for depression OR depressive OR mental OR psychiatric OR anxiety OR PTSD OR phobia OR OCD AND feedback.

There were no restrictions on date, language or publication status applied to the searches.

4. **Update searches 2015**

An update search was performed on 18 May 2015 to identify additional RCTs eligible for inclusion. At this time we thought it appropriate to validate the 2014 searches by checking the (a) the provenance of included studies (to date) and (b) information contained in the title,abstract and subject heading fields of study reports in MEDLINE, EMBASE and PsycINFO. This exercise revealed that eight of the eleven included studies (>70%) were only identified from screening reference lists or from the Web of Science citation search and four of these studies made no mention of the patients' mental health condition. The searches were overhauled and the PsycINFO and CCDANCTR databases re-searched all years to 18 May 2015, together with a search of the Cochrane Library (Appendix 2). A further citation search of WoS was also conducted, to 27 May 2015.

5. **Update searches 2016**

In compliance with MECIR conduct standard 37 we ran an update search within 12 months of publication (on 25 May 2016), including the following databases: PsycINFO, CCDANCTR, CENTRAL, Web of Science, and the ICTRP/ClinicalTrials.gov international trial registries. These results have not yet been incorporated into the review.

Searching other resources

Grey literature

Google Scholar (top 100 hits) and Google.com were searched (verbatim) for: "Patient Reported Outcome Measures" and "mental health" and (randomised or randomized). Search results were screened for relevant reports and reviews.

Reference lists and correspondence

We screened reference lists (of trial reports and systematic reviews) to identify additional studies missed from the original electronic searches (including unpublished or in-press citations); used the related articles feature in PubMed; and contacted other experts and trialists in the field for information on unpublished or ongoing studies, or to request additional trial data. 'Patient reported outcome measures' and 'PROMs' are relatively recently adopted terms in the literature. For earlier studies, where the terminology used may be ambiguous, we had to rely more on these informal methods of discovery.

Data collection and analysis

Selection of studies

Two review authors (TK and ME-G) independently screened titles and abstracts for inclusion of all the potential studies identified as a result of the search, coded as 'retrieve' (eligible or potentially eligible or unclear) or 'do not retrieve'. We resolved disagreements through discussion and consultation with a third author (MM). We retrieved the full-text study reports or publications and the same two review authors independently screened the full texts, identified studies for inclusion, and identified and recorded reasons for exclusion of the ineligible studies. Again, disagreements were resolved through discussion and consultation with the third author MM. We excluded duplicate records and collated multiple reports that related to the same study so that each study rather than each report became the unit of interest in the review. We recorded the selection process in sufficient detail to complete a PRISMA flow diagram (Moher 2009) and 'Characteristics of excluded studies' table.

Data extraction and management

We designed and used a data collection form which was piloted on one study in the review to extract study characteristics and outcome data. Five review authors (TK, ME-G, AB, LA, ALB) independently extracted study characteristics and outcome data from the included studies. We extracted the following study characteristics.

1. Methods: study design (cluster or individual randomisation), total duration of study, number of study centres and location, study setting, withdrawals, and dates of study.
2. Participants: n, mean age, age range, gender, severity of condition, diagnostic criteria (clinical only, DSM or ICD, etc.), inclusion criteria, exclusion criteria, and co-morbidities.
3. Interventions: intervention including the specific instrument(s) used and whether the results were fed back to the treating clinician only or also to the participant; whether education about interpretation and an algorithm were also provided; and details of treatment as usual provided to the comparison group.
4. Outcomes: primary and secondary outcomes specified and collected, and time points reported.
5. Notes: funding for trial, and notable conflicts of interest of trial authors.

We noted in the 'Characteristics of included studies' table if outcome data were not reported in a usable way. We resolved disagreements by consensus and also by involving a third person (MM). Two review authors (TK, ME-G) transferred data into Review Manager (RevMan) (RevMan 2014), and double-checked that data were entered correctly by comparing the data presented in the systematic review with the study reports. Another two review authors (BS, AG) spot checked the accuracy of data extracted, against the original study reports.

Main comparison

1. Treatment informed by feedback of patient reported outcome measures compared with treatment as usual.

Assessment of risk of bias in included studies

Two review authors (TK and ME-G) independently assessed the risk of bias for each study using the criteria outlined in the *Cochrane Handbook for Systematic Reviews of Interventions* (Higgins 2011). We resolved any disagreements by discussion and by involving other authors (MM, BS, RC, SG). We assessed the risk of bias according to the following domains:

1. Random sequence generation.
2. Allocation concealment.
3. Blinding of participants and clinicians (performance bias, which will be high due to the nature of the intervention).
4. Blinding of researchers conducting outcome assessments (detection bias).
5. Incomplete outcome data.
6. Selective outcome reporting.
7. Other bias.

We judged each potential source of bias as high, low or unclear and provided supporting quotations from the study report where available, together with a justification for our judgment in the 'Risk of bias' table. We summarised the risk of bias judgements across different studies for each of the domains listed. We considered blinding separately for different key outcomes where necessary. Where information on risk of bias related to correspondence with a trialist, we noted this in the 'Risk of bias' table.

When considering treatment effects, we took into account the risk of bias for the studies that contributed to that outcome.

Measures of treatment effect

Continuous data

We calculated mean differences (MD) and the associated 95% confidence interval (CI) for continuous outcomes where there was a common measure across studies, and standardised mean differences (SMD) and the associated 95% CI where different scales were used to measure the same underlying construct. We entered the data presented as a scale with a consistent direction of effect.

Dichotomous data

We carried out a narrative analysis to describe categorical outcomes. See Differences between protocol and review.

Unit of analysis issues

Cluster randomised trials

Clustering by clinician, clinic, practice or service would be the preferred design over randomising individual participants since a clustered design reduces the risk of contamination between arms, as the PROMs are not routinely available in the control settings and are therefore much less likely to be used inadvertently in control patients. However, failure to account for intra-class correlation in clustered studies is commonly encountered in primary research and leads to a 'unit of analysis' error (Divine 1992) whereby P values are spuriously low, CIs unduly narrow, and statistical significance overestimated, causing type I errors (Bland 1997; Gulliford 1999). For studies that employed a cluster randomisation, we sought evidence that clustering was accounted for by the authors in their analyses.

Studies with multiple treatment groups

Where multiple trial arms were reported in a single trial, we included all relevant arms that compared treatment as usual with routine outcome monitoring.

Where we found three-armed trials that compared PROMs fed back to the clinician only, versus PROMs fed back to both the clinician and patient, versus treatment as usual, we divided the control group between the two comparisons so as not to use the same data twice, which would constitute a unit of analysis error. However, we also performed a sensitivity analysis excluding any trials with this three-arm design from the subgroup analysis (see below) to see whether this significantly affected the results of the subgroup analysis.

Dealing with missing data

We contacted investigators in order to verify key study characteristics and obtain missing numerical outcome data, where possible. We documented all correspondence with trialists and report which trialists responded below. (If standard deviations were missing, we planned to calculate them, if possible, from the available information reported (including 95% CIs and P values) or impute standard deviations from similar studies using the same instruments, but in the event we did not need to do this).

Assessment of heterogeneity

Between-study heterogeneity was assessed using the I^2 statistic (Higgins 2003), which describes the percentage of total variation across studies that is due to heterogeneity rather than chance. A rough guide to interpretation is as follows: 0% to 40% might not be important; 30% to 60% may represent moderate heterogeneity; 50% to 90% may represent substantial heterogeneity; and 75% to 100% considerable heterogeneity. We investigated the sources of heterogeneity as described below where the I^2 value was greater than 50%. Where I^2 was below 50% but the direction and magnitude of treatment effects suggested important heterogeneity, we also investigated the potential sources.

Assessment of reporting biases

We created funnel plots where feasible and where there were sufficient studies (that is 10) (Egger 1997) to investigate possible publication bias. Funnel plot tests for asymmetry were separately conducted in STATA (StataCorp. 2015), using the metabias command.

Data synthesis

We undertook meta-analyses only where it was meaningful, that is where the PROM feedback interventions, participants and the underlying clinical question were similar enough for pooling to make sense. We pooled change scores as a first preference where these were available, checking assumptions about the approximate normality of data by ensuring that the difference between the mean and lowest or highest possible value divided by the standard deviation was greater than two. Less than two would indicate some skew and less than one would indicate substantial skew. We planned not to attempt pooling for data that were substantially skewed and where the skew could not be reduced by transforming the data. We planned to describe skewed data as medians and interquartile ranges.

We anticipated significant heterogeneity between studies (I^2 value of over 50%) as we were including a range of CMHDs, a range of settings, and both self-complete and administered outcome measures. Therefore we used a random-effects model when combining data to minimise the effect of heterogeneity between studies. Where studies were combined which used outcome measures that scored treatment effects in opposite directions, the mean values of one set of studies were multiplied by -1 to ensure the scales identified benefited in the same direction, in accordance with section 9.2.3.2 of the *Cochrane Handbook* (Deeks 2011).

Where cost data were presented and a formal cost-effectiveness analysis had been undertaken, we planned simply to describe the

methods and results. We did not plan to attempt formal statistical pooling of cost data because studies often adopt different perspectives; account for different types of cost data; use different methods of discounting future healthcare costs and benefits; are conducted at different points in time; and are conducted in different countries with varying funding and reimbursement systems, making international comparisons difficult.

Subgroup analysis and investigation of heterogeneity

We planned to conduct the following six subgroup analyses, which should be regarded as exploratory since they are observational and not based on randomised comparisons. We planned to restrict these six subgroup analyses to the three primary outcomes (namely improvement in symptom scores, health-related quality of life, and adverse effects).

1. Whether the setting of the study (primary care, multidisciplinary mental health services, or psychological therapies) influenced the success of the strategy.
2. Studies in which a formal diagnosis (according to DSM or ICD criteria) was made prior to treatment using a validated assessment, versus studies of participants diagnosed on clinical assessment only, as the formally diagnosed group were likely to be more homogeneous and more alike in their responses to PROMs.
3. Studies of participants aged 18 to 65 years versus those with participants aged over 65 years, as the older age group may have more complex disorders with co-morbid cognitive changes and it is plausible that recovery follows a different pathway.
4. Studies where feedback of the results of PROMs was given only to the clinician versus studies where feedback was given to both clinician and participant, as the previous review by Knaup 2009 showed a greater effect when patients were also given feedback.
5. Studies where feedback of the results of PROMs was given only to the participating patient versus studies where feedback was given to the clinician only, or to both clinician and patient, if any such studies were identified (we thought this was unlikely given the results of previous systematic reviews of outcome monitoring in mental health, which have not identified any studies of feedback to patients alone).
6. Studies where feedback to the clinician included treatment instructions or an algorithm for actions to be taken for particular results, compared to studies where feedback was limited to the results of the PROM alone, to determine whether treatment recommendations in addition to PROM results influenced the results.

Post-hoc subgroup analyses

We decided post-hoc to conduct an additional subgroup analysis, comparing studies involving Michael Lambert, the originator and owner of the OQ-45 system, with studies not involving him, to explore whether potential benefits of the system were identified in independent evaluations. This was because the OQ-45 was the PROM used in the large majority of studies in the meta-analyses, and Michael Lambert was author or co-author of a significant proportion of those studies (see section on Differences between protocol and review).

We also decided during the course of the review to meta-analyse results for subgroups of participants within studies who were identified as being at higher or lower risk for treatment failure, which was determined by the trajectory of their initial response to therapy. The low risk group was described as 'on-track' (OT) for a good clinical response, and the high risk group as 'not on track' (NOT). This was a post-hoc change to the methods which we agreed due to the fact that several identified studies reported potentially important findings in analyses of outcomes for subgroups of OT and NOT participants. One comparison included only the NOT subgroup, comparing outcomes in terms of symptom scores between feedback and non-feedback arms. The second comparison included both the OT and NOT subgroups, comparing the number of treatment sessions received between feedback and non-feedback arms, and including a formal test for subgroup differences to look for evidence of differences between OT and NOT subgroups. This was a further change from the protocol, as the number of treatment sessions was a secondary outcome, and originally we planned to conduct subgroup analyses restricted to the three primary outcomes, namely symptoms, health-related quality of life, and adverse effects (see section on Differences between protocol and review).

Sensitivity analysis

We planned to conduct the following sensitivity analyses to explore their effects on the results obtained in the review, and to test the robustness of decisions made in the review process:

1. Whether the mode of administration (self-complete versus clinician-rated) influenced the success of the strategy, by re-analysing after removing studies using clinician-rated PROMs and seeing whether the result was significantly different.
2. Whether cluster randomised studies produced a different result from non-clustered studies, to see whether possible contamination between arms in non-clustered designs reduced the difference between arms, by re-analysing after removing non-clustered studies.
3. Within cluster RCTs, whether adjustment for unit of analysis error influenced the results, to test the robustness of the results arising from non-adjusted analyses.
4. Whether the inclusion of quasi-randomised cluster trials significantly affected the results, by re-analysing after removing quasi-randomised cluster trials.
5. Whether losing the data from three-arm trials that compared PROMS fed back to the clinician only, versus PROMS fed back to both the clinician and patient, versus treatment as usual, made a significant difference to the results of

the subgroup analysis (4 above), by excluding such trials from the subgroup analysis.

'Summary of findings' tables

We developed 'Summary of findings' tables to summarise the key findings of the review, for the populations in primary care, multidisciplinary mental health care, and psychological therapy settings. We tabulated the comparisons between PROMs and usual care in terms of effects on participant outcomes including symptoms, social functioning, quality of life and adverse effects; and on the process of care including drug prescriptions and referrals. Decisions on which measurements to incorporate into the 'Summary of findings' table were based on those most relevant to clinical practice, taking into consideration the specific nature of the scale and also the time points at which measurements were made. We used the GRADE criteria to assess the body of evidence for each comparison.

RESULTS

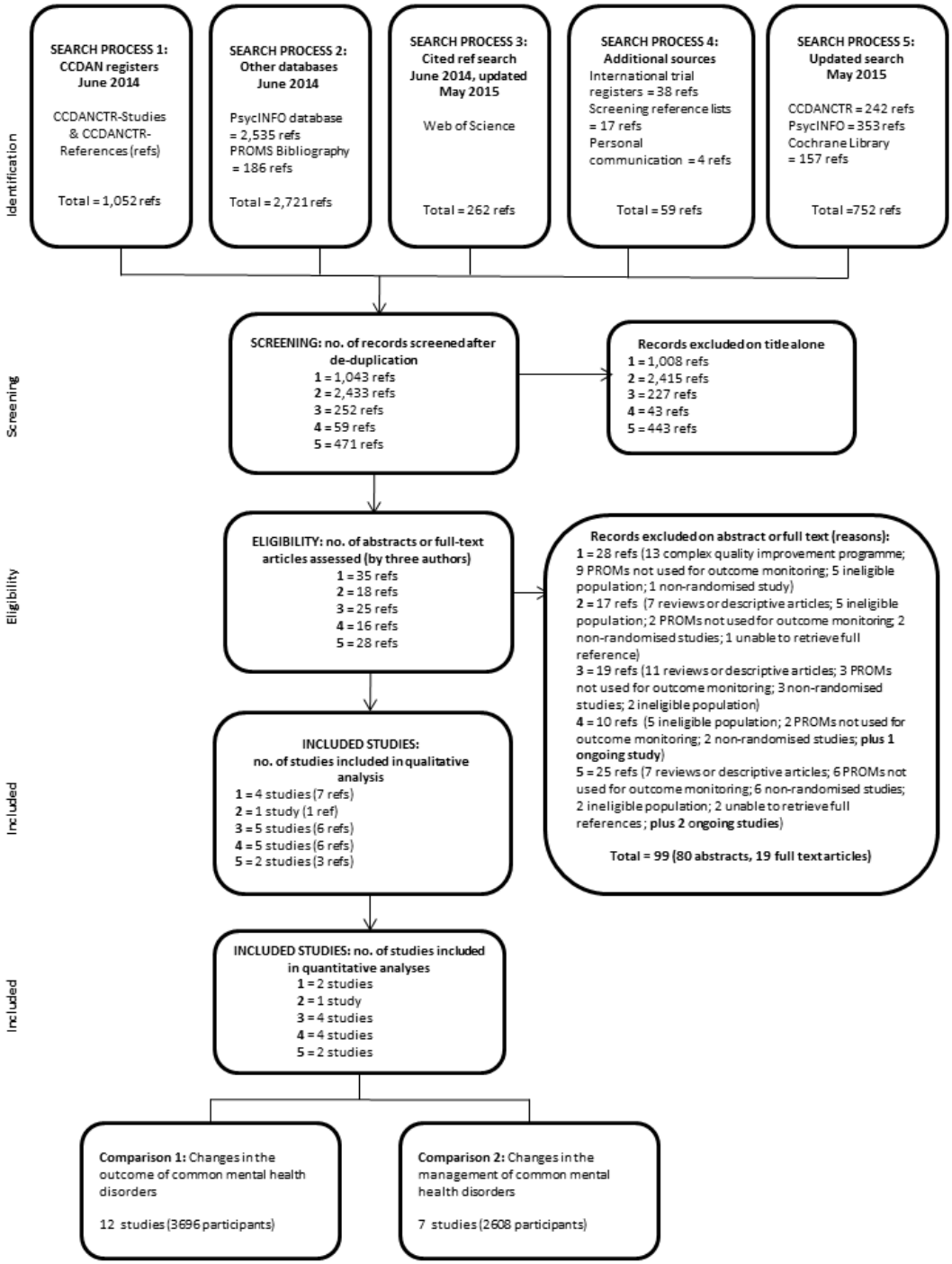
Description of studies

Seventeen studies met our inclusion criteria: Amble 2014; Berking 2006; Chang 2012; De Jong 2012; De Jong 2014; Hansson 2013; Hawkins 2004; Lambert 2001; Mathias 1994; Murphy 2012; Probst 2013; Reese 2009a; Reese 2009b; Scheidt 2012; Simon 2012; Trudeau 2001; and Whipple 2003

Results of the search

The initial searches of CCDANCTR, OVID PsycINFO and PROM bibliographies (to 30 May 2014) yielded 1052, 2535, and 186 references respectively (see PRISMA diagram, Figure 1). The WoS citation search to 5 June 2014 yielded 262 references, and we identified a further 59 references through searching the international trial registers, screening reference lists, and personal communication with trial authors. An updated search (to 18 May 2015) was conducted to validate identified references by re-searching PsycINFO and CCDANCTR along with *The Cochrane Library*, which yielded a further 752 references. Following de-duplication, we screened a total of 4258 references obtained through these searches, of which we excluded 4136 on assessment of the title alone. Of the remaining 122, 99 were excluded on the basis of reading and discussing the abstract (80) or full-text article (19), including 25 reviews or descriptive articles, 22 where PROMs were not used for outcome monitoring, 19 with ineligible populations (adolescents, severe mental illness, eating disorders, or substance misuse), 14 non-randomised studies, 13 which included complex quality improvement programmes, three because we were unable to retrieve full references, and three ongoing studies (NCT01796223; NCT02023736; NCT02095457); see PRISMA diagram (Figure 1). Further information is given below on the 19 studies excluded on the basis of reading and discussing the full-text articles (see Excluded studies).

Figure 1. PRISMA flow diagram



In compliance with MECIR conduct standard 37 we ran an update search within 12 months of publication (on 25 May 2016), including the following databases: PsycINFO (which identified 72 references), CCDANCTR (29 references), CENTRAL (37), Web of Science (139), and the ICTRP/ClinicalTrials.gov international trial registries (28): in total 305, and de-duplicated 281 references. This update search identified two additional completed studies (Gibbons 2015 and Rise 2016) which are awaiting classification, and four additional ongoing studies (Metz 2015; NCT02656641; NTR5466; and NTR5707). These results will be fully incorporated into the review at the next update (as appropriate).

The remaining 23 references described 17 included studies, of which 13 (Amble 2014; De Jong 2012; De Jong 2014; Hansson 2013; Hawkins 2004; Lambert 2001; Murphy 2012; Probst 2013; Reese 2009a; Reese 2009b; Simon 2012; Trudeau 2001; and Whipple 2003) were included in quantitative meta-analyses as they used comparable outcome measures (either the Outcome Questionnaire (OQ-45, Lambert 2004) or Outcome Rating System (ORS, Miller 2003), see interventions below), and the remaining four (Berking 2006; Chang 2012; Mathias 1994; Scheidt 2012) were included in the qualitative assessment (see PRISMA flow diagram, Figure 1).

The results of attempts to clarify study details through contacting authors are given in the table below. Contact details were unobtainable for the authors of Mathias 1994. Of those contacted seven authors responded (with regard to De Jong 2012; De Jong 2014; Haderlie 2012; Hansson 2013; Hawkins 2004; Puschner 2009; Reese 2009a; Reese 2009b; and Trudeau 2001;), and the remainder failed to respond (with regard to Chang 2012; Lambert 2001; Probst 2013; Simon 2012; and Whipple 2003).

Included studies

The individual studies are described in detail in the Characteristics of included studies table below.

Design

Thirteen studies were randomised at the individual level and four were cluster randomised (Chang 2012; Mathias 1994; Reese 2009b; Scheidt 2012). Fourteen studies had one intervention arm in which feedback of patient reported outcomes was given, and one control arm in which patients completed the measures but the results were not fed back. De Jong 2014; Hawkins 2004; and Trudeau 2001 included three arms: De Jong 2014 and Hawkins 2004 included two intervention arms, one in which feedback was given to the clinician only and one where feedback was given to both clinician and patient; and Trudeau 2001 included an additional control arm in which patients were not asked to complete the measures at all.

Sample sizes

The number of participants per study ranged from 96 to 1629 with a total of 8787 participants. A substantial proportion of participants were not used in data analysis due to withdrawal or loss to follow-up, with all but two studies (De Jong 2012; Hansson 2013) utilising only a per protocol analysis. This number totaled 2650 (30.1%).

Setting

The majority of the studies (nine) were carried out in the USA. The remainder were carried out in Germany (three), The Netherlands (two), Sweden (one), Norway (one) and Ireland (one). Fifteen studies were conducted exclusively in outpatient settings, and two, Berking 2006 and Probst 2013, were inpatient studies. One study (Amble 2014) included both inpatients and outpatient clinics. Seven studies were multi-centre with the remainder confined to one site.

Two studies were based in primary care settings (Chang 2012; Mathias 1994); nine in multidisciplinary mental health care settings (Amble 2014; Berking 2006; De Jong 2012; De Jong 2014; Hansson 2013; Hawkins 2004; Probst 2013; Simon 2012; Trudeau 2001); and six in psychological therapy settings (Lambert 2001; Murphy 2012; Reese 2009a; Reese 2009b; Scheidt 2012; Whipple 2003).

Participants

The 17 included studies comprised 8787 randomised participants (pre-attrition total), of whom 6137 (69.9%) provided follow-up data and were included in the study analyses. The age of participants ranged between 18 to 75 years, but in several studies the range was not reported. The median age across the studies was 35.1 years. The proportion of women among participants ranged from 58% to 73%, although there was inconsistency in reporting, with some studies providing the proportion of women among participants randomised, and some the proportion among participants included in the analysis. Reporting of demographic details was quite variable between studies, with marital status and employment being the most commonly recorded demographics. In studies which reported on ethnicity, the majority of participants were white.

Fourteen studies reported specific diagnoses for their participants, of which three used ICD diagnostic criteria (Amble 2014; Berking 2006; Scheidt 2012), and three used DSM criteria (De Jong 2012; De Jong 2014; Mathias 1994). The remaining studies characterised participants on the basis of clinical diagnoses rather than diagnostic criteria. Three studies did not report the specific diagnoses of their participants (Reese 2009a; Reese 2009b; Trudeau 2001),

and five did not assign a specific diagnosis of a CMHD to 20% or of their participants, reporting that they had interpersonal or relationship difficulties, other diagnoses including personality or behavioural disorders, or were given administrative codes (Amble 2014; De Jong 2014; Lambert 2001; Murphy 2012; Whipple 2003).

Interventions

Feedback was usually given in the form of scores on the PROMs, together with information on whether this meant the participant had improved or not. Feedback was given only to the clinician in six studies: Chang 2012; Hawkins 2004 (one arm); Mathias 1994; Probst 2013; Scheidt 2012; and Trudeau 2001. Feedback was given explicitly to both the clinician and participant in seven: De Jong 2014 (one arm); Hansson 2013; Hawkins 2004 (one arm); Murphy 2012; Reese 2009a; Reese 2009b; and Simon 2012. In the other seven studies clinicians were permitted or encouraged to share feedback with the participant: Amble 2014; Berking 2006; De Jong 2012; De Jong 2014 (one arm); Lambert 2001; Probst 2013; and Whipple 2003.

Eight different PROMs were used across the studies, the most common being the Outcome Questionnaire-45 (OQ-45, Lambert 2004), a compound measure of psychiatric symptoms, individual functioning, interpersonal relations, and performance in social roles, which was used in 10 studies (Amble 2014; De Jong 2012; De Jong 2014; Hansson 2013; Hawkins 2004; Lambert 2001; Probst 2013; Simon 2012; Trudeau 2001; Whipple 2003). As well as the OQ-45 scores, feedback was colour coded to allow quick appreciation of the extent of change during a busy clinic. In three of these studies (Probst 2013; Simon 2012; Whipple 2003) additional interventions were applied in the 'not on-track' (NOT) groups, giving clinicians specific instructions on whether or not to change treatment according to the results of the outcome measure, and what further treatments to apply, known as the 'Assessment of Signal Cases' (ASC), and 'Clinical Support Tool' (CST) respectively.

Three studies (Murphy 2012; Reese 2009a; and Reese 2009b) used a shorter measure derived from the OQ-45, known as the Outcome Rating System (ORS, Miller 2003) which includes the same domains as the OQ-45.

The duration of the treatment period was variable, being determined by the clinician or patient terminating treatment in most studies, and so the duration of follow-up was also variable, as the final measure of outcome was usually collected at the last treatment session.

Outcomes

Our primary outcome (mean change in symptom score) was reported by all studies, but of the remaining two primary outcomes health-related quality of life was assessed by only two of the trials (Mathias 1994; Scheidt 2012), and adverse effects (including suicide and self-harm) were also assessed by only one (Chang 2012). Changes in the management of the CMHD (pharmacological treatment and referral to secondary care) were reported by two studies (Chang 2012; Mathias 1994), and eight studies reported effects on the number of treatment sessions received by participants (Amble 2014; De Jong 2014; Hawkins 2004; Lambert 2001; Reese 2009a; Reese 2009b; Simon 2012; Whipple 2003).

Timing of outcome assessment

All but two of the studies reported research outcomes only in the short-term, up to six months after baseline assessment. De Jong 2014 and Scheidt 2012 also reported longer-term outcomes, after 35 weeks and 12 months respectively.

'On track' and 'not on track' participants

In 10 studies (De Jong 2012; De Jong 2014; Hansson 2013; Hawkins 2004; Lambert 2001; Murphy 2012; Reese 2009a; Reese 2009b; Simon 2012; Whipple 2003) results were reported for subgroups of participants according to whether they were identified early in their treatment as 'on-track' (OT) or 'not on track' (NOT) for a good clinical response. The NOT group were also sometimes labelled as 'at risk', 'signal cases', or 'signal alert cases'.

Excluded studies

After obtaining and assessing the full text of the report we excluded 19 studies. Six studies were non-randomised, six did not use the PROM for outcome monitoring or did not report patient outcomes, five included an ineligible population, and two involved the use of a PROM as part of a more complex quality improvement programme. See Characteristics of excluded studies for further details.

Ongoing studies

We identified seven ongoing studies that fitted our inclusion criteria. Three of these studies are comparing feedback to the therapist only with treatment as usual (TAU) (NCT01796223; NCT02023736 and NCT02095457); two are comparing feedback to both therapist and participant with TAU (Metz 2015 and NTR5466); one is comparing feedback to the participant only with TAU (NTR5707), and one has a three-arm design comparing feedback to the participant only, versus feedback to the participant and therapist, versus TAU (NCT02656641). Outcome is being measured with the OQ-45 and Beck Depression Inventory (BDI) in NCT02023736; with the CORE-OM rating scale (Barkham 2006) in NCT02095457 (together with health service utilisation and level of functioning); with the Outcome Rating Scale (ORS) in NCT01796223 and NTR5466; with the

PHQ-9 and GAD-7 in NCT02656641; and with the Inventory of Depressive Symptomatology (IDS) plus the OQ-45 in NTR5707. The primary outcome in Metz 2015 is the enablement of shared decision making measured using the Decisional Conflict Scale (DCS), but the OQ-45 outcome measure is one of the secondary outcomes. Four studies are taking place in multidisciplinary mental health care settings (Metz 2015; NCT01796223; NCT02095457; NTR5707), two in psychological therapy settings (NCT02023736; NCT02656641), and one in both primary care and multidisciplinary mental health care settings (NTR5466). See Characteristics of ongoing studies for further details.

Studies awaiting classification

We identified two completed and published studies, Gibbons 2015 and Rise 2016, which are awaiting classification.

Gibbons 2015 cluster randomised people with depression attending a community mental health centre in Philadelphia, USA, to eight weeks of individual therapy with either a clinician receiving weekly feedback reports, or a clinician not receiving weekly feedback reports, using the BASIS-24 (24-Item Behavior and Symptom Identification Scale) as both a PROM and outcome measure. The study reported a medium effect size in favour of the feedback condition for symptom improvement (Effect size (Cohen's d) = 0.50, P value = 0.017), and 36% of feedback participants compared to 13% of participants in the no feedback condition demonstrated clinically significant change across treatment (P value = 0.013).

Rise 2016 cluster randomised outpatients attending a mental health hospital in Norway to feed back to both client and therapist using the PCOMS system PROMs (Outcome Rating Scale (ORS) and Session Rating Scale (SRS)), or to care without feedback, measuring outcomes with the Behaviour and Symptom Identification Scale 32 (BASIS-32), and Patient Activation Measure (PAM). The study reported that, at 6 and 12 months after starting treatment there were no significant effects on the primary outcomes of mental health symptoms or patient activation, but compared to baseline assessment the PCOMS group had significantly improved their patient activation scores after 12 months.

See Characteristics of studies awaiting classification for further details.

Risk of bias in included studies

We categorised the overall risk of bias for each study, with all studies considered at high risk of bias (a plausible presence of bias that seriously weakens confidence in the results), as one or more domains received a judgement of high risk. In most studies, we judged inadequate blinding and attrition at high risk of bias, see sections below for further details. For details of the risk of bias judgement for each study, see Characteristics of included studies. A graphical representation of the overall risk of bias of included studies is presented in Figure 2 and Figure 3.

Figure 2. Risk of bias summary: review authors' judgements about each risk of bias item for each included study.

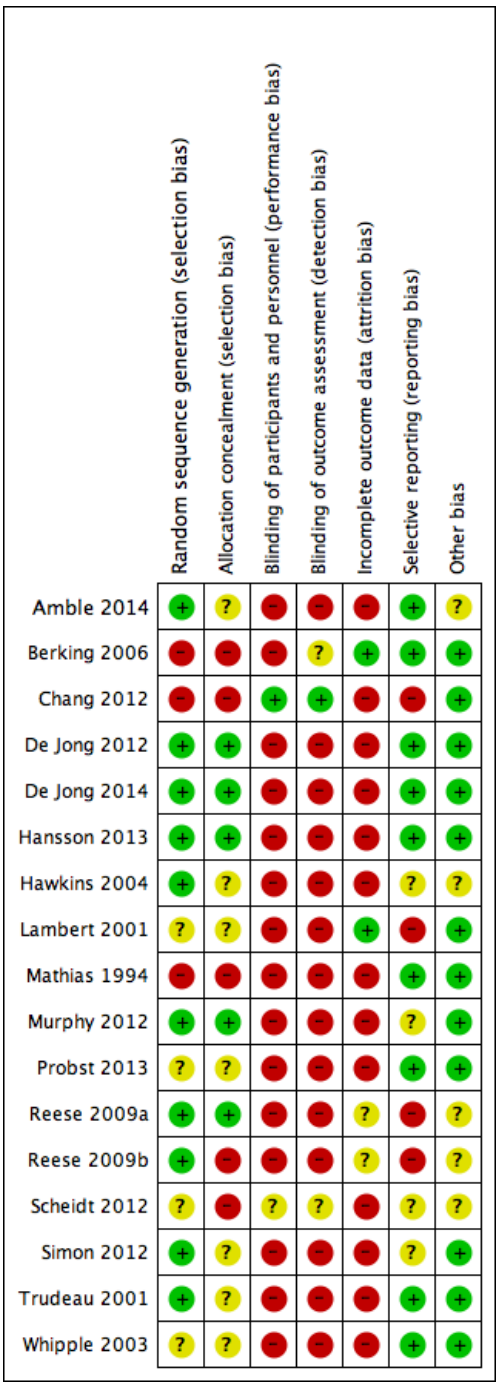
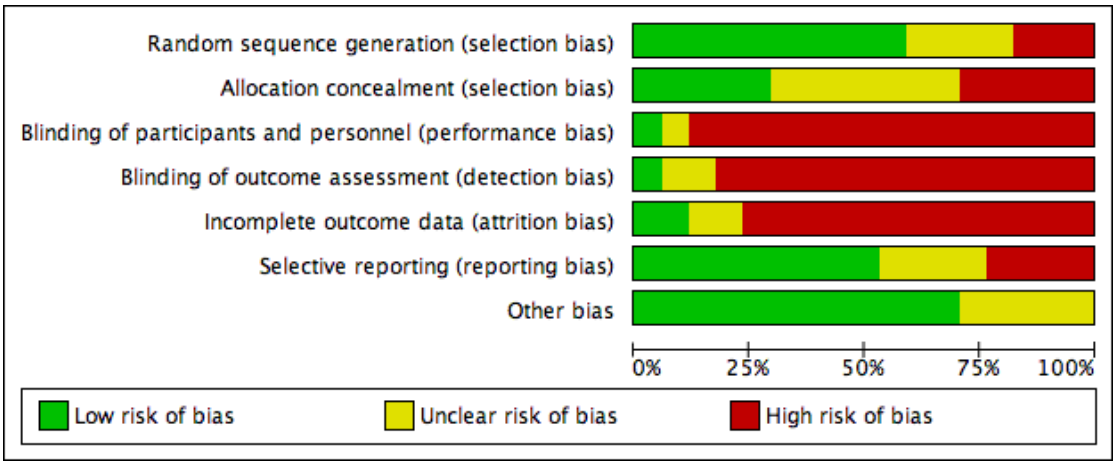


Figure 3. Risk of bias graph: review authors' judgements about each risk of bias item presented as percentages across all included studies.



Allocation

This review included studies using either an individual randomisation or cluster randomisation method. The studies that were cluster randomised may be, by method, at risk of selection bias, where concealment of group allocation prior to obtaining consent is not ensured, thereby increasing the likelihood that participants allocated to the intervention group have a propensity to using PROMs. This bias risk is not commented upon by any of those studies using a cluster randomisation, and therefore, it is not possible to exclude a high risk of selection bias.

Sequence generation

Seven of the studies described the means to generate the allocation sequence in sufficient detail to make a favourable assessment of whether comparable groups should be produced (Amble 2014; De Jong 2014; Hansson 2013; Hawkins 2004; Murphy 2012; Simon 2012; Trudeau 2001). After contact with the authors, Reese 2009a; Reese 2009b; and De Jong 2012 were also deemed to have used an adequate means of sequence generation, therefore in total 10 studies were judged at low risk of bias for this domain. A lack of adequate reporting in four studies (Lambert 2001; Probst 2013; Scheidt 2012; Whipple 2003) resulted in a judgement of unclear risk of bias. The remaining three studies were judged at high risk of bias for this domain, as Berking 2006 used coin tossing, and Chang 2012 and Mathias 1994 assigned clinics and call centres, respectively, to feedback and no-feedback groups without details of how they were randomised.

Allocation concealment

Insufficient details were reported regarding allocation concealment in nine studies, rendering a judgement of unclear risk (Amble 2014; Hawkins 2004; Lambert 2001; Probst 2013; Reese 2009a ; Scheidt 2012; Simon 2012; Trudeau 2001; Whipple 2003). Four studies were considered at low risk of bias (De Jong 2012; De Jong 2014; Hansson 2013; Murphy 2012), while Chang 2012; Mathias 1994; Reese 2009b; Scheidt 2012 were considered to be at high risk, due to their cluster randomised design.

Blinding

Due to the nature of the intervention, it is very difficult to blind the clinicians in studies fitting the inclusion criteria for this review. However, we judged the majority of the studies at high risk of bias as the group allocation was clearly known to the participating clinicians in all but two studies. Chang 2012 was considered to be at low risk as participants in both arms received feedback but the frequency of feedback varied between the two arms, and the participants were unaware of which arm they were in. In the intervention arm feedback was monthly through the six-month study period and could therefore influence outcome, while in the control arm feedback was not provided until the end of the six-month

period, and so could not affect outcome. Insufficient details were reported in Scheidt 2012, so the risk of bias for this study was judged as unclear.

Similarly, the risk of bias related to blinding of outcome assessors was judged as high for all but three of the 17 studies. In 13 studies the PROM used for feedback was also used for outcome assessment, so the participants themselves were the outcome assessors and they were not blind to whether or not they received the intervention. In the remaining three, the risk of bias was judged to be high in one (Mathias 1994) as the researchers assessing outcome were apparently aware of group allocation; unclear in two, as it was not reported whether allocation was concealed from the outcome assessors (Berking 2006; Scheidt 2012); and low in one (Chang 2012), as, although the PROM used for feedback was also used for outcome assessment, the participants receiving the feedback were unaware of which arm they were in.

Incomplete outcome data

The attrition rate, through loss to follow-up, was considered to put study results at high risk of bias in all but two of the studies, Berking 2006 and Lambert 2001. This was usually because participants were excluded from the analysis if they did not have at least two outcome measures completed (and sometimes three), as the change in outcome from baseline was measured by the clinician at the last therapy session, and there was no measure of outcome outside the clinical setting by an independent researcher at a predetermined follow-up point. This may be justified in the sense that only participants who have at least a second outcome measure completed might be considered to have had the 'minimum dose' of feedback necessary to examine its effects, but it means that all of the studies except two were analysed on a per protocol basis rather than according to intention to treat (ITT). Hansson 2013 did report an ITT analysis, and De Jong 2012 used multiple imputation to deal with the problem of missing data at follow-up.

Lambert 2001 reported no drop-outs, with all 609 study participants recruited apparently completing the study, but we were unable to confirm that with the author. The risk of incomplete outcome data was low for Berking 2006 as that study of inpatients collected follow-up data on more than 98% of the participants.

Selective reporting

We looked for published protocols for the included studies, in order to determine whether selective reporting had taken place, but were unable to identify any. Judged from the aims and methods described in the included study reports, the risk of selective reporting bias was judged to be low for nine studies (Amble 2014; Berking 2006; De Jong 2012; De Jong 2014; Hansson 2013; Mathias 1994; Probst 2013; Trudeau 2001; Whipple 2003), while four were judged to have a high risk of reporting bias due to incomplete reporting of primary outcomes (Chang 2012; Lambert 2001; Reese 2009a; Reese 2009b), and for four (Hawkins 2004, Murphy 2012, Scheidt 2012; Simon 2012) it was unclear whether selective reporting had taken place.

Ten studies using the OQ-45 or ORS PROMs reported results separately for 'on track' (OT) and 'not on track' (NOT) subgroups of participants, but the criteria for defining these subgroups were specified using a priori definitions in the OQ-45 and ORS systems, and the results were usually reported for both OT and NOT subgroups, except for Simon 2012 which did not report data for OT participants.

Other potential sources of bias

We did not identify any other sources of bias.

Effects of interventions

See: **Summary of findings for the main comparison** Treatment informed by feedback of patient reported outcome measures compared with treatment as usual

Comparison 1: Treatment informed by feedback of patient reported outcome measures compared with treatment as usual

Primary Outcomes

1.1 Mean improvement in symptom scores

Our primary analysis compared feedback to clinician (with or without feedback to the participant in addition), to no feedback (usual care). Pooling of patient outcome data across studies to provide a summary estimate of effect was possible only for those studies measuring outcome using the OQ-45 or ORS compound outcome measures as PROMs. There was no evidence of skew, with all values being greater than 2.

a) OQ-45 PROM

Nine studies (Amble 2014; De Jong 2012; De Jong 2014; Hansson 2013; Hawkins 2004; Lambert 2001; Probst 2013; Trudeau 2001; Whipple 2003) including 3438 participants contributed data to a comparison of studies using the OQ-45 as both PROM and outcome measure. (Simon 2012 also used the OQ-45 but could not be included as the study reported results only for 'not on track' patients, see post-hoc analysis below). This analysis revealed no evidence of a difference between feedback and no-feedback groups in terms of symptom scores (mean difference (MD) -1.14, 95% CI -3.15 to 0.86; P = 0.26, I^2 = 25%), see Analysis 1.1. The evidence for this comparison was considered low quality. All but one of these studies reported outcomes only in the short-term, up to 26 weeks post-baseline. De Jong 2014 reported no significant differences

overall between outcomes for short-term (up to 35 weeks) and long-term (35 to 78 weeks) treatment.

b) OQ-45 and ORS PROMs

We combined an additional three studies (Murphy 2012; Reese 2009a; Reese 2009b), with a further 258 participants using the ORS as both a PROM and outcome measure with the studies above in Analysis 1.2 (see Summary of findings for the main comparison). Again, this analysis revealed no evidence of a difference between feedback and no-feedback groups in terms of symptom scores (standardised mean difference (SMD) -0.07 95% CI-0.16 to 0.01; P = 0.10, I^2 = 30%). The evidence for this comparison was also considered low quality.

c) Other outcome measures

Four studies identified used a variety of measures of global symptoms, depressive symptoms alone, anxiety symptoms alone, or quality of life alone, which we judged too dissimilar to combine, in terms of the domains measured, with the results for the OQ-45 or ORS, which are compound outcome measures combining symptoms, functioning and relationships, and so these four studies were not included in the meta-analyses.

Two studies in US primary care populations reported mixed findings.

Chang 2012 measured depressive symptoms using the PHQ-9 depression questionnaire as both PROM and outcome measure, and reported a significantly greater odds of response in terms of changes in scores on the PHQ-9 significant difference between feedback (n = 364) and no feedback (n = 278) groups (odds ratio (OR) 2.02, 95% CI 1.36 to 3.02).

Mathias 1994 developed a composite PROM, the Mental Health Patient Profile, constructed from the symptom checklist SCL-90, diagnostic interview schedule DIS, and quality of life short form SF-36 scales. Outcome was measured in terms of overall symptom severity using the Global Severity Index (GSI), and anxiety symptoms using the Highest Anxiety Subscale Score (HASS). They reported no significant differences between feedback (n = 367) and no feedback (n = 216) groups in either GSI or HASS. At completion of the study, the mean GSI for the feedback group was 59.90 and for no feedback 60.89, P = 0.89. Similarly, no significant difference between feedback and no-feedback groups was seen in mean HASS scores (feedback group 64.72, no feedback 68.23, P = 0.74).

The quality of evidence for these outcomes was graded as moderate.

Two German studies used compound questionnaires as both PROMs and outcome measures.

Berking 2006 used a compound questionnaire for assessing success and course of psychotherapeutic treatment (FEV) to monitor progress in inpatients in multidisciplinary mental health care. The FEV included an emotionality inventory (EMI-B); a brief symptom inventory (BSI); an inventory of interpersonal problems (IIP-D); and a measure of cognitive changes (INK), and gave a compound score. FEV scores were compared between feedback (n = 40) and no-feedback groups (n = 39) at the end of their inpatient stay, and a more favourable effect was reported in the feedback group: a change from a mean pre-study score of 2.90 (SD 0.62) to 2.25 (0.71), compared with 2.91 (0.69) to 2.54 (0.77) in the no feedback group.

Scheidt 2012 also used a compound comprehensive inventory of psychometric measurement instruments to monitor psychotherapy outpatients. Based on the scores, decision rules ('reorientation of the expert system') were developed and optimised to guide decisions about indications for, and prolongation of, psychotherapy based on the feedback received ('TK system'). The compound measure also assessed several outcomes: a brief symptom inventory (BSI), inventory of interpersonal problems (IIP-D), Beck Depression Inventory (BDI); and a questionnaire on body-related anxiety and cognitions (AKV). There was no difference seen in the BSI between the feedback (302) and no-feedback groups (160) at 12 months post treatment (MD 1.00, 95% CI -2.22 to 4.22; P = 0.54). There was a slightly better outcome seen in the IIP-D score in the feedback group (n = 305, control n = 158) but this was not significant (MD 2.30, 95% -0.37 to 4.97; P = 0.09). However, the feedback group (n = 205) scored significantly better on the BDI questionnaire than the no feedback group (n = 124) (MD 4.60, 95% CI 0.79 to 8.41; P = 0.02). On assessment with the AKV, there was no difference seen between groups (feedback n = 71, control n = 24) (MD -1.50, 95% C.I. -7.31 to 4.31; P = 0.61).

The quality of evidence for these outcomes was graded as moderate to low.

1.2 Health-related quality of life

We did not pool data for this outcome as it was reported by only two studies.

Mathias 1994 assessed quality of life using the 36-item Short Form (SF-36) scale. They reported no significant difference between feedback (n = 367) and no feedback (n = 216) groups in any of the nine sub scales of the SF-36. Mean mental sub scale scores were 66.0 for feedback and 64.8 for no-feedback groups (P = 0.31).

Scheidt 2012 assessed quality of life using the 12-item Short Form (SF-12) scale, reporting results for mental and physical sub scales. There were no significant differences seen between feedback (n = 376) and no feedback (n = 211) groups at the end of treatment, for both physical (MD -0.90, 95% CI -3.11 to 1.31; P = 0.55) and mental sub scales (MD 1.20, 95% CI -0.51 to 2.91; P = 0.55).

The quality of evidence for these comparisons was graded as moderate to low.

1.3 Adverse events

Only one study reported any findings in relation to adverse events: in Chang 2012, 273 PHQ-9 questionnaires elicited thoughts of suicide or self-harm and no immediate suicide risk was discerned. However, information on which arms these findings were in was not provided.

Adverse events from prescribed medication were not assessed in any of the included studies.

Secondary outcomes

1.4 Changes in management of CMHDs

a) Changes in prescribed drug treatment

Only two studies reported differences in changes of prescribed drug treatment. Chang 2012 showed that, at six months, 200/352 in the feedback and 115/252 in the no feedback group had no change in pharmacological treatment. Percentages without a change in antidepressant therapy did not differ significantly between study arms (OR 1.21 95% CI 0.78 to 1.88; P = 0.06). Mathias 1994 also reported no significant difference between feedback and no-feedback groups in changes in prescriptions for psychotropic medications (OR 1.09, 95% CI 0.94 to 1.85).

b) Referrals

Only Mathias 1994 assessed levels of referral to 'mental health specialists', without distinguishing between psychiatry and psychology referrals, and reported that referrals were significantly more likely in the feedback group (OR 1.73, 95% CI 1.11 to 2.70).

c) Number of treatment sessions received

An addition to the planned comparisons of changes in management of CMHDs was made post-hoc (see Differences between protocol and review section), namely an analysis of differences in the mean number of treatment sessions received between feedback and no-feedback groups. Data from seven studies that reported numbers of treatment session (Amble 2014; De Jong 2014; Hawkins 2004; Lambert 2001; Reese 2009a; Reese 2009b; Whipple 2003) were pooled, including 2608 participants, in an analysis which showed no evidence of a difference in the mean number of treatment sessions between feedback and no-feedback groups (MD -0.02 sessions, 95% CI -0.42 to 0.39; P = 0.93, I^2 = 0%), see Analysis 1.3 and Summary of findings for the main comparison.The quality of evidence for this outcome was graded as low.

Probst 2013 also reported no significant differences between feedback and no-feedback groups in the number of weeks (rather than sessions) of treatment received.

1.5 Social functioning

Only one study reported differences in social functioning. Hansson 2013 found no differences between feedback and no-feedback groups in mean scores on the sub scale of the OQ-45 relating to social functioning (feedback group (n = 136) 13.9, no feedback group (n = 126) 14.9, P = 0.10).

1.6 Costs

No studies reported any cost data.

Subgroup analyses

2. Whether the setting of the study influenced the success of the strategy

We could only carry out subgroup comparisons for the primary outcome of mean improvement in symptom scores, due to the lack of data on health-related quality of life and adverse effects.

2.1 Setting 1: Primary care

2.1.1 Mean improvement in symptom scores

Neither study from a primary care setting was included in the meta-analyses above, as they did not use the OQ-45 or ORS, and so could not be considered within this subgroup analysis.

2.2 Setting 2: Multi-disciplinary mental health care settings

2.2.1 Mean improvement in symptom scores

Seven studies in multidisciplinary mental health care (Amble 2014; De Jong 2012; De Jong 2014; Hansson 2013; Hawkins 2004; Probst 2013; Trudeau 2001) comprising 1848 participants assessed the effect of the OQ-45 questionnaire as a PROM feedback tool. These data were pooled in a meta analysis, see Analysis 2.1.1 This showed no evidence of a difference between feedback and no-feedback groups (SMD -0.05, 95% CI -0.18 to 0.07; P = 0.40, I^2 = 37%). The quality of evidence for this comparison was also graded as low.

2.3 Setting 3: Psychological therapies

2.3.1 Mean improvement in symptom scores

Five studies undertaken in psychological therapy service settings (Lambert 2001; Murphy 2012; Reese 2009a; Reese 2009b; Whipple 2003) comprising 1848 participants were pooled in a meta-analysis assessing the effect of the OQ-45 or ORS as a feedback tool, see Analysis 2.1.2. This again showed no evidence of a difference between feedback and no-feedback groups in terms of symptom scores (SMD -0.10, 95% CI -0.23 to 0.03; P = 0.14, I^2 = 29%). The quality of evidence for this comparison was also graded as low.

There was no significant difference between the results obtained for the subgroup of studies carried out in multi-disciplinary mental

health care settings and those carried out in psychological therapy settings (test for subgroup differences: Chi² = 0.23, df = 1 (P = 0.63), I² = 0%).

3: Whether participants who had a formal diagnosis made using ICD or DSM criteria were likely to do better than those where no formal diagnosis was made

3.1 Mean improvement in symptom scores

In three studies within the meta-analysis (Amble 2014; De Jong 2012; De Jong 2014), including 1144 participants, the participants were given formal diagnoses. There was no evidence of a difference between feedback and no-feedback groups (SMD -0.01, 95% CI -0.23 to 0.21), see Analysis 3.1.1 and no evidence of benefit was observed in the subgroup analysis limited to the nine studies in which no formal diagnoses were given (SMD -0.08, 95% CI -0.15 to 0.00; P = 0.06, I^2 = 0%), see Analysis 3.1.2. There was no significant difference between the subgroup of studies where a formal diagnosis was given and those without a formal diagnosis (test for subgroup differences: Chi² = 0.28, df = 1 (P = 0.60), I² = 0%). The quality of evidence for these outcomes was graded as low.

4: Studies of participants aged 18 to 65 years versus those with participants aged over 65 years

This planned subgroup comparison was not possible as no studies distinguished subgroups of participants in the two age categories.

5: Studies where feedback was given only to clinicians versus studies where feedback was given to both clinicians and participants

5.1: Mean improvements in symptom scores

Feedback was given only to the clinician in six studies: Chang 2012; Hawkins 2004 (one arm); Mathias 1994; Probst 2013; Scheidt 2012; and Trudeau 2001. Feedback was given explicitly to both the clinician and patient in seven: De Jong 2014 (one arm); Hansson 2013; Hawkins 2004 (one arm); Murphy 2012; Reese 2009a; Reese 2009b; and Simon 2012. In the other seven studies clinicians were permitted or encouraged to share feedback with the patient: Amble 2014; Berking 2006; De Jong 2012; De Jong 2014 (one arm); Lambert 2001; Probst 2013; and Whipple 2003. Subgroup analyses showed no evidence of differences in outcomes between these three groups as the confidence intervals overlapped for all analyses (a formal statistical test for difference between subgroups was not carried out as some studies contributed data to more than one subgroup, potentially violating the assumption of independence). See Analysis 4.1.

6: Studies where feedback of PROM results was given only to the participants and not to the clinicians

This planned subgroup analysis was not possible as no studies were identified in which feedback was given only to participants and not to clinicians.

7: Studies where feedback to the clinician included treatment instructions or an algorithm in addition to a score on a PROM

7.1 Mean improvement in symptom scores

Two studies in the meta-analysis, including 1184 participants, included treatment instructions in the form of a clinical support tool (CST) for those found to be not on track (NOT): Probst 2013 and Whipple 2003. (Simon 2012 also included a CST but provided data only for the NOT participants). A subgroup analysis comparing these two with the remainder showed no significant differences in either subgroup (test for subgroup differences: Chi² = 0.64, df = 1 (P = 0.42), I² = 0%). For those studies with a CST: SMD -0.03 (95% CI -0.14 to 0.09; P = 0.66, I^2 = 0%), compared to: SMD -0.09 (95% CI -0.20 to 0.02; P = 0.11, I^2 = 38%) for those studies without a CST, see Analysis 5.1.1 and 5.1.2.

Post-hoc sub-group analysis of studies involving Michael Lambert, the originator of the OQ-45 PROM and feedback system, versus studies not involving him

8.1 Mean improvement in symptom scores

Six studies included in this review involved Michael Lambert as either first author or co-author. There was no significant difference in the overall findings in terms of outcomes observed between a subgroup of five studies which included him as an author (see Analysis 6.1.1) and four which did not (Analysis 6.1.2) (test for subgroup differences Chi² = 0.18, df = 1 (P = 0.67), I² = 0%).

Post-hoc analyses of subgroups of 'on track' and 'not on track' participants

1. Improvement in symptom scores among 'not-on-track' participants

Ten studies including 923 participants identified participants who were considered 'not on track' (NOT), 'at risk', or 'signal alert cases' early on during their treatment, and provided separate data for these participants, see Analysis 7.1. Symptom scores were slightly

lower in the feedback group compared to the no feedback group in this subgroup (SMD = -0.22, 95% CI -0.35 to -0.09; P = 0.001, I^2 = 0%). The quality of evidence for this comparison was graded as low.

2. Number of treatment sessions received: 'on track' and 'not on track' participants

Five studies reported differences in the number of treatment sessions received between feedback and no-feedback groups for NOT participants. Data from De Jong 2014; Hawkins 2004; Lambert 2001; Reese 2009b and Whipple 2003 were pooled in a meta-analysis which demonstrated no evidence of a difference in the mean number of therapy sessions received (see Analysis 7.2). In addition, Probst 2013 reported that for the NOT subgroup of participants, there was no significant difference between feedback and no-feedback groups in the number of weeks (rather than sessions) of treatment received: 6.22 (SD 3.29) compared to 5.49 (3.17), P = 0.46. Four studies (De Jong 2014; Lambert 2001; Reese 2009b; Whipple 2003) reported differences in the amount of therapy received between feedback and no-feedback groups for the subgroup of 'on track' (OT) participants only. The mean number of treatment sessions was slightly fewer in the feedback group: (MD -0.69, 95% CI -1.10 to -0.29; P = 0.0007, I^2 = 0%). However, a formal test for subgroup differences revealed no significant difference between findings for OT and NOT participants (chi^2 = 0.99, df = 1 (P = 0.32), I^2 = 0%, see Analysis 7.2). The quality of evidence for this comparison was also graded as low.

Sensitivity analyses

The following planned sensitivity analyses were not possible:

1. Whether the mode of administration (self-complete versus clinician-rated) influenced the success of the strategy, because the main analysis did not include any studies using clinician-rated PROMs.
2. Whether cluster randomised studies produced a different result from non-clustered studies, because the main analysis included only one cluster randomised study, Reese 2009b.
3. Within cluster RCTs, whether adjustment for unit of analysis error influenced the results, again because the main analysis included only one cluster randomised study, and Reese 2009b did report that the results were adjusted for clustering.
4. Whether the inclusion of quasi-randomised cluster trials significantly affected the results, because the main analysis did not include any quasi-randomised cluster trials.

The only sensitivity analysis which could be carried out was:

5. Whether losing the data from three-arm trials (that compared PROMs fed back to the clinician only, versus PROMs fed back to both the clinician and participant, versus treatment as usual), made a significant difference to the results of the subgroup analysis, by excluding such trials from the subgroup analysis.

Two studies (De Jong 2014; Hawkins 2004) included the three arms. No difference was seen in the overall result when these studies were omitted altogether from the meta-analysis: (SMD -0.02, 95% CI -0.011 to 0.07; P = 0.68, I^2 = 39%), although the direction of the treatment effect was slightly more in favour of the feedback group. Excluding the two studies which included three arms did not make a difference to the subgroup analysis which still showed no significant difference between the subgroups (test for subgroup differences: Chi^2 = 1.92, df = 2; P = 0.38, I^2 = 0%).

Post-hoc sensitivity analysis: unreported or incomplete diagnoses of study populations

Three studies included in the meta-analyses did not report the specific diagnoses of their participants (Trudeau 2001; Reese 2009a; Reese 2009b), and five did not assign a specific diagnosis of a CMHD to 20% or more of their participants (Lambert 2001; Whipple 2003; Murphy 2012; De Jong 2014; Amble 2014). A sensitivity analysis of the meta-analysis of studies using the OQ-45 (Analysis 1.1) omitting Lambert 2001; Trudeau 2001; Whipple 2003; De Jong 2014 and Amble 2014 showed no evidence of a difference in the overall result (MD -0.94, 95% CI -3.67 to 1.78). Similarly, a sensitivity analysis of the meta-analysis of studies using either the OQ-45 or ORS (Analysis 1.2), omitting all eight studies, showed no evidence of a difference (SMD -0.03, 95% CI -0.12 to 0.06).

Exploration of heterogeneity

We found I^2 values between 30% and 69%, indicating moderate to significant heterogeneity, in our main meta-analysis of the difference in outcome feeding back OQ-45 or ORS scores versus no feedback, and in seven of our subgroup analyses. Investigating the sources of heterogeneity in these comparisons, we found that they nearly all included Amble 2014 which reported distinctly positive findings. The authors reported that only 25% of the therapists employed at their clinics agreed to participate, which is a small proportion compared to the other studies included, which usually involved all or most of the therapists working in a service. Also, three quarters of the participants included were seen in the clinic where the project leader and main coordinator worked, underlining "the importance of having a dedicated local advocate monitoring and following up the procedures for using a feedback system" (Amble 2014, p.6). It seems likely therefore that the therapists seeing clients in that study were self-selected for their enthusiasm for routine monitoring.

Reporting Bias

Funnel plots conducted for publication bias in relation to the meta-analyses of outcomes measured using the OQ-45 only (Analysis 1.1) and OQ-45 plus ORS (Analysis 1.2), are shown in Figure 4 and Figure 5 respectively. The Egger test (Egger 1997) indicated that there was no evidence of publication bias in Analysis 1.1 (P = 0.499) or Analysis 1.2 (P = 0.512).

Figure 4. Funnel plot of comparison: 1 Difference in outcome feeding back OQ-45 or ORS scores versus no feedback, outcome: 1.1 Mean improvement in symptom scores: OQ-45 PROMS.

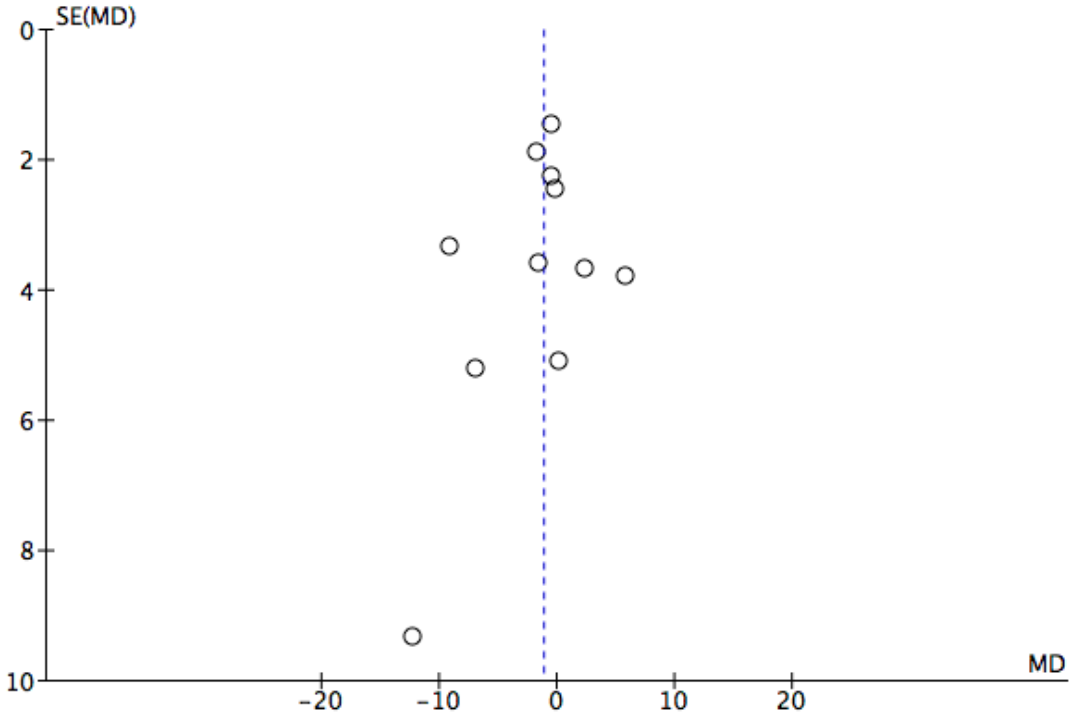
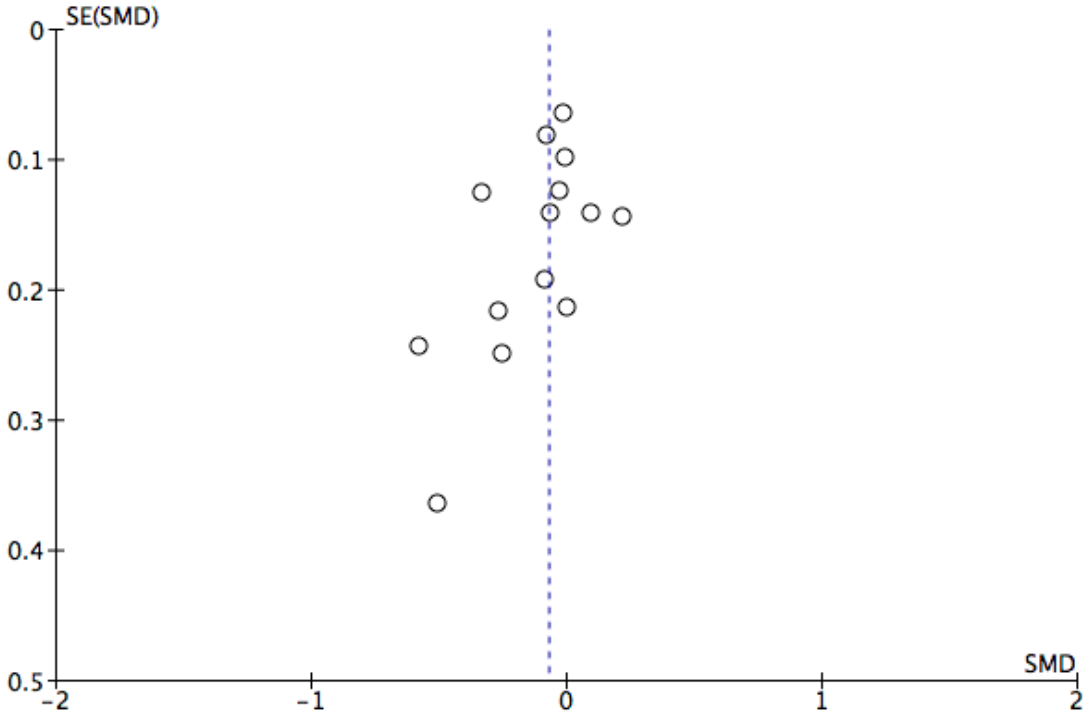


Figure 5. Funnel plot of comparison: 1 Difference in outcome feeding back OQ-45 or ORS scores versus no feedback, outcome: 1.2 Mean improvement in symptom scores: OQ-45 or ORS PROMs.



DISCUSSION

Summary of main results

In terms of improvements in the outcome of CMHDs, we found no evidence of a difference between feedback and no-feedback groups in our meta-analysis of 12 studies using the OQ-45 or ORS PROMs. We also found no evidence of an effect on the management of CMHDs in terms of the number of treatment sessions participants received, in an analysis combining the results of seven studies using the OQ-45 or ORS (see Summary of findings for the main comparison). However, because the evidence we identified is of low quality, we are uncertain about this result, and further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate. The majority of the eligible studies we identified were conducted in multidisciplinary mental health care settings (nine) or psychological therapy settings (six). We identified only two eligible studies conducted in primary care settings, and we were unable to include them in our meta-analyses as they measured research outcomes with single symptom, global improvement, or quality-of-life domains, which were different to the OQ-45 and ORS which are compound outcome measures combining symptoms, social functioning and interpersonal relationships. A qualitative assessment of the two primary care studies showed conflicting findings: Chang 2012 found a significant difference in outcome but Mathias 1994 found none, while Mathias 1994 found significant effects on the management of CMHDs but Chang 2012 found none (see Summary of findings for the main comparison).

We did find a difference in outcomes favouring feedback, but with a small effect size (a standardised mean difference of -0.22), in a post-hoc meta-analysis including only the subgroup of 'not on track' (NOT) participants monitored in 10 studies using the OQ-45 or ORS PROMs. We also found a small reduction in the number of treatment sessions received for the 'on-track' (OT) subgroup in another post-hoc analysis of four studies using the OQ-45 or ORS (a mean difference of -0.69 sessions), but a formal test of subgroup differences revealed no evidence of a difference in the number of treatment sessions received between OT and NOT subgroups of patients.

Overall completeness and applicability of evidence

Most of the outcomes that we anticipated might be affected were not addressed by the majority of studies. Health-related quality of life, social functioning, and adverse events were each considered only in one or two studies, while the costs of intervention were not considered in any (see Summary of findings for the main comparison). Similarly, most studies did not report changes in the management of CMHDs in terms of drug treatment or referrals to other specialities, and only a minority reported effects on the number of sessions of treatment received by participants. We were therefore unable to answer all the questions we set out to address. The meta-analyses we conducted included studies in multidisciplinary mental health care and psychological therapy settings only, with no inclusion of studies in primary care, as both primary care studies we identified used outcome measures with different domains which were not directly comparable to the OQ-45 or ORS.The evidence we have been able to analyse is therefore largely limited to the effects of compound PROMs measuring psychological symptoms, functioning, and relationships in one instrument rather than single domain instruments, and in multidisciplinary mental health care and psychological therapy settings, rather than in primary care. This is unfortunate given that the large majority of people with CMHDs are treated in primary care, if they receive treatment at all (McManus 2009), at least in countries with well developed systems of primary care. This contrasts with the relatively large number of studies which have been conducted using patient reported measures for screening and initial identification of CMHDs in primary care (Gilbody 2008). Our findings are therefore consistent with those of Shaw 2013, who concluded that there is a lack of evidence to support recommendations for routine monitoring of people with CMHDs using PROMs in primary care.

It is perhaps not surprising that most of the studies have been conducted in psychological therapy or multidisciplinary mental health care settings, given that people attending these services have a relatively more homogeneous set of presenting problems, and the service staff have a relatively homogeneous professional background, because the services are limited to mental health problems. Introducing routine outcome monitoring of CMHDs in primary care is more challenging, given that primary care deals with the whole range of initially undifferentiated physical, mental and social problems, and so only a minority of people seen in that setting will have CMHDs. In services dedicated to psychological therapies and mental health it is likely to be easier to train staff to routinely administer PROMs to all attending patients, and develop efficient administrative systems and information technology to support monitoring, whereas in primary care staff firstly have to decide which patients have CMHDs and then whether to administer PROMs when relevant, and developing systems to deal with routine outcome monitoring may be regarded as less worthwhile when the patients to be monitored are in a minority. It is questionable therefore whether any benefits identified from routine outcome monitoring in psychological therapy or mental health care settings can be extrapolated to primary care, as patient engagement, the development of routine systems for administering PROMs, and the technical resources and healthcare professional training required to interpret them, may all be more challenging in the primary care setting where only a proportion of patients present with CMHDs.

Quality of the evidence

We rated the quality of the evidence, summarised in Summary of findings for the main comparison, as low or moderate. The main reasons for the low quality rating were limitations in study design, with regard to lack of blinding of clinicians, participants and outcome assessors; attrition; and indirectness of the evidence. All of the included studies were judged to be at high risk of bias and so we exercise caution in our interpretation of the findings based on the low to moderate quality observed. Specific considerations are discussed below.

Limitations in study design and implementation

The quality of evidence in almost all studies was downgraded due to issues of blinding. Chang 2012 was the only study that executed a study design that blinded clinicians and participants to the differences in frequency and timing of feedback between intervention and control arms, and this study was not included in any of the meta-analysis. Due to the absence of such blinding, most studies were considered high risk with respect to this domain (see Figure 2). Additionally, all but three studies were judged at high risk of bias when considering the blinding of outcome assessors, as in most studies the PROMs used for feedback were also used for outcome assessment. Shimokawa 2010 called this the "monomethod", and it increases the risk of observer rating bias, as the therapist using the PROM as a clinical intervention can potentially influence the research outcome when it is based on the same PROM. Only one study, Chang 2012 was judged to be at low risk of bias for outcome assessment as, although the PROM used for feedback was also used for outcome assessment, the study was cluster randomised and participants receiving the feedback were not aware of the two conditions operating.

Problems contacting several of the study authors meant no clarification could be obtained regarding randomisation and blinding, so we could not upgrade the evidence for these studies.

A significant proportion of the participants were lost to follow-up, around 30% altogether, and in all but two studies a per protocol analysis was carried out. Only two studies, Berking 2006 and Lambert 2001, were graded at low risk of bias as there was no attrition at all reported in these two studies. Per protocol analyses were usually conducted because the PROM was used not only as the clinical intervention but also to measure the research outcome, and there was no independent follow-up by researchers after participants had completed therapy.

Consistency of effect

The effect of treatment across the studies was fairly consistent, with minimal heterogeneity observed in the main analyses. The effect of feeding back PROMs was modest and present in most studies and any inconsistency was minimal within a small and overall insignificant treatment effect. Heterogeneity was more apparent in those subgroup analyses which included Amble 2014, which was unusual in demonstrating an apparently high level of clinician commitment to using PROMs.

Imprecision of the results

The evidence was not downgraded for imprecision in Summary of findings for the main comparison as the sample sizes exceeded the optimal information size with resultant narrow confidence intervals. However, due to the limitations of study design and implementation, we have been cautious in interpreting the precision of results overall. Precision was reduced in the analysis of subgroups of OT and NOT participants although this is to be expected due to the reduced sample size for each subgroup.

Indirectness of the results

The review included studies involving several groups of participants: primary care patients, clients attending for psychological therapies, and clients under the care of multidisciplinary healthcare teams, both as inpatients and outpatients. A broad spectrum of participants was therefore included, but only two studies were based in primary care, and pooling of data was not possible for these studies as they used PROMs which did not measure the same domains as those used in psychological therapy and multidisciplinary mental health care settings. The OQ-45 compound outcome measure was used in 10 of the included studies and the ORS, which was derived from the OQ-45, was used in another three. The preponderance of studies using these scoring systems means we have minimal evidence on which to base judgements of the use of other, quicker to administer, single domain PROMs such as the PHQ-9 which is widely used in the USA (HRSA 2005) and UK (IAPT 2011).

All included studies provided outcome data on the change in symptom scores, but it was disappointing that most of the other outcomes pre-specified for this review were not considered by the trials identified. Numbers of treatment sessions were reported in half of the studies, but other important indicators of changes in management such as drug prescription changes or referral for further treatment were reported in only two, and consideration of adverse events, social functioning, and estimates of costs, were almost completely lacking. Consequently, the quality of evidence has to be rated as low in relation to the indirectness of the results.

Publication bias

The funnel plots (Figure 4; Figure 5) and Egger tests for publication bias suggest publication bias was not an issue in this review. However it must be noted that there were only just sufficient numbers of studies to be considered for a formal publication bias assessment.

Potential biases in the review process

We carried out a comprehensive search for eligible studies, using multiple electronic databases, followed up with searching of reference lists, citation searches, and contact with study authors who identified further studies in some cases, in particular Kim de Jong who was therefore invited to be a co-author. We were also able to include two German studies, as we recruited Anna Brütt, who is German, as a co-author to help with data extraction. However, all but two of the studies identified were published after the year 2000, apart from Brody 1990 and Mathias 1994, which gives cause for concern that we might have missed studies published in the period 1994-2000. Although thorough and comprehensive searches were performed to identify all potential studies for inclusion, the searches were initially very inefficient due to the rather non-specific terms 'feedback' and 'monitor*' capturing many papers which were not actually about using feedback as an intervention to monitor patients' progress. This became more apparent on running the citation searches which yielded further study reports which were not originally identified. We were also unable to contact study authors in a number of cases, in particular Michael Lambert who was a co-author on six studies, who might have been able to identify further studies for us. It is possible therefore that we failed to identify relevant studies.

A number of studies did not characterise a significant proportion of their participants in terms of underlying diagnoses, as indicated in the Characteristics of included studies section, and in three studies no diagnoses were reported at all (although the lead authors confirmed the large majority participants had CMHDs). We intended to include samples without a formal diagnosis, as they are common especially in the psychological therapy setting, but we also made post-hoc decisions to include studies where 20% or more of participants were reported to have relationship or interpersonal problems, or received administrative codes only, as long as the majority of those diagnosed were given a diagnosis of a CMHD, and as long as fewer than 10% were diagnosed with a severe mental illness, substance misuse, learning difficulty, dementia, or eating disorder. The lack of specific diagnoses for many participants is a significant limitation of the available literature, and we recommended that future studies characterise the diagnoses of all their participants systematically.

Agreements and disagreements with other studies or reviews

Our results are less positive in terms of favouring the routine use of PROMs than those reported in an earlier meta-analysis of three studies in psychological therapy settings using the OQ-45 (Lambert 2003), which reported a small but significant overall positive effect on outcome (an effect size of 0.09), and a larger significant positive effect in the subgroup of NOT participants (effect size of 0.39, compared to 0.22 in our analysis). However, Lambert 2003 included a study (Lambert 2002) which we judged non-randomised as it used 'historical controls' (i.e. it used archived data from clients previously treated in the clinic as control data, rather than randomising subjects to a control arm) and was therefore excluded from our analysis, as non-randomised studies confer a greater risk of bias.

Our findings also differ from those of Knaup 2009, who reported an overall significant positive effect of routine monitoring on outcomes (effect size 0.10) from a meta-analysis of 12 studies in multidisciplinary mental health care and psychological therapy settings. Knaup 2009 also reported a greater effect size (0.30) for studies including feedback of PROM results to patients than for those where feedback was given only to the clinician (0.09), which we did not find. However, Knaup 2009 included only five of the studies included in our review (Berking 2006; Hawkins 2004; Lambert 2001; Trudeau 2001; Whipple 2003), and a further seven studies which were excluded from this study, including two non-randomised studies (Lambert 2002; Slade 2008), and five which were conducted with people with eating disorders or severe mental illness. Again, non-randomised studies confer a greater risk of bias, and studies including people with severe mental illness might be more positive, as patients' symptoms are more severe, so the potential for improvement is greater than among people with CMHDs, where a possible 'floor effect' might limit the potential to show a benefit from monitoring with PROMs.

More positive findings than ours were also found by Shimokawa 2010 in an update of the Lambert 2003 meta-analysis, which added three more studies using the OQ-45, and reported an overall effect size of 0.12 in favour of outcome monitoring. They also reported a larger effect size among the NOT subgroup of participants, of 0.28, in an intention to treat (ITT) analysis utilising last observations carried forward (LOCF), and an even larger effect size among NOT participants in a per protocol analysis, of 0.53. However, three of the six studies included in Shimokawa 2010 were not randomised trials (Harmon 2007; Lambert 2002; Slade 2008), as they used 'historical controls' and so again were excluded from our analyses due to the increased risk of bias. We were unable to carry out any corresponding ITT analyses using LOCF data, as we could not obtain further data from study authors for several of the studies we included.

Another possible reason for differences between our review and the earlier reviews is that ours included later studies which did not involve the originators of the OQ-45 system, whose authors might therefore have had less allegiance to the system, and less adherence to its founding principles, thereby diluting its effects. However, we found no evidence of differences between intervention and control groups in a post-hoc subgroup analysis of the results of five studies authored or co-authored by Michael Lambert, the originator of the OQ-45 system, which was similar to the findings among four studies which did not involve him.

Our findings are consistent with those of a more recent systematic review of studies limited to the use of the OQ-45 or ORS in psychological therapy settings (Davidson 2014), which also concluded that the benefit of feedback monitoring appeared to be limited to NOT participants, although Davidson 2014 did not conduct a meta-analysis. Their review also included the three studies with historical controls (Harmon 2007; Lambert 2002; Slade 2008), as well as a study of people with substance misuse (Crits-Christoph 2012), and one of people with eating disorders (Simon 2013), all of which were ineligible for inclusion in this review. Davidson 2014 pointed out, as we have, that many studies were of low quality due to methodological issues.

Our findings relating to the primary care setting are also consistent with those of Shaw 2013, the main finding being a distinct lack of research on the monitoring of CMHDs with PROMs in primary care when compared with multidisciplinary mental health care and psychological therapy settings. We were able to identify only two trials (one of which, Chang 2012, was considered by Shaw 2013), which reported conflicting findings in terms of impacts on the outcome and management of CMHDs. Our findings in primary care settings are also consistent with Gilbody 2002 who failed to identify a positive impact of patient-centred outcome instruments assessing patient needs or quality of life in non-psychiatric settings.

Our findings are less positive than those of Carlier 2012 and Poston 2010 which both included studies of the use of PROMs as screening or diagnostic tools together with studies of their use as follow-up monitoring measures, and so were not directly comparable. They are more consistent with Boyce 2013, Marshall 2006 and Valdera 2008, who all found the evidence of benefit from monitoring with PROMs to be weak, although again they are not directly comparable to our review, as they included studies of the use of PROMs in the management of physical disorders as well as studies in mental health care. Carlier 2012 recommended further research was needed in mental health care, a recommendation we make below. Boyce 2013 and Valdera 2008 pointed out that most of the studies they identified suffered from methodological limitations, as we have found in this review, and that there was significant heterogeneity.

Heterogeneity in this review was apparently related to clinician commitment to using PROMs, as the outstandingly positive finding was found by Amble 2014 in which therapists self-selected as participants for their interest in using the OQ-45 PROM. De Jong 2012 looked at therapist variables that moderated feedback effects, and found that improved outcomes in NOT patients were associated with greater commitment to using feedback, perceived validity of feedback, and self-efficacy among participating therapists.

Other factors which have been suggested by Krageloh 2015 as important in explaining differences in the findings between trials include: having a formalised structure which maximises the likelihood that feedback is discussed with clients; the use of computerised support tools; greater frequency of feedback; and whether PROMs are discussed with clinicians, although Krageloh 2015 did not conduct any meta-analyses to support those suggestions. We found no difference in outcome between our subgroup analysis of two studies where a clinical support tool (CST) was used to guide responses to scores on the OQ-45, compared to the remaining studies without CSTs. It has also been suggested that feedback given to both clinicians and patients is more effective than feedback to clinicians alone (De Jong 2012; Hawkins 2004; Knaup 2009), but we did not find that to be the case in our analyses of three subgroups: feedback limited to the clinician; feedback which could be shared with the patient; and feedback routinely provided to the patient as well as the clinician.

Our findings may be contrasted with those of a Cochrane review of collaborative care for depression and anxiety disorders, which found that, compared to usual care, it was associated with significant improvement in symptoms, quality of life, and patient satisfaction (Archer 2012). Collaborative care usually includes feeding back the results of PROMs at initial assessment and follow-up to inform treatment, but collaborative care includes a number of other active components such as medication management and increased liaison between healthcare professionals, and the process of measuring and feeding back patient outcomes was actually the control condition in some trials of collaborative care interventions (Archer 2012).

AUTHORS' CONCLUSIONS

Implications for practice

On the basis of this review, no firm conclusions can be drawn about the effects of routine monitoring of patients with common mental health disorders using patient reported outcome measures.

The meta-analyses including all participants monitored with the OQ-45 or ORS PROMs across both psychological therapy and multidisciplinary mental health care settings found very small differences between the feedback and no-feedback groups in terms of outcome, which may not be clinically meaningful (see below). There was no difference in management in terms of the number of treatment sessions received by participants overall.

The two studies conducted in primary care which were not included in the meta-analyses showed conflicting findings in terms of both outcome and changes in management, and the lack of studies conducted in primary care means no conclusions can be drawn about the likely value of routine outcome monitoring in that setting.

This review therefore provides little support for policy recommendations in the UK and USA that people with CMHDs should be routinely monitored using PROMs (HRSA 2005; IAPT 2011; NICE 2011b). In particular we agree with Davidson 2014 that the available trial evidence, coming largely from the USA and Europe, and mostly using the compound OQ-45 or ORS outcome measures, has limited generalisability to the IAPT psychological therapy settings in the UK, where routine outcome monitoring of hundreds of thousands of people with CMHDs takes place every year using a range of single domain PROMs (HSCIC 2015). Our findings are also consistent with those of Shaw 2013, who concluded there is a lack of evidence to support recommendations for routine monitoring of people with CMHDs with PROMs in primary care.

The low quality of evidence found means we are uncertain about these results however, and further research is very likely to have an important impact on our confidence in the estimate of effect, and is likely to change the estimate.

Implications for research

More trials of routine outcome monitoring in CMHDs using PROMs are needed, particularly in primary care settings in the UK and elsewhere, where most people with CMHDs are treated, and should include more people treated with antidepressants as well as those treated with psychological therapies, since antidepressants are the commonest treatments for CMHDs provided in primary care (Kendrick 2009). PROMs which have fewer items, such as the ORS or PHQ-9, may be preferable due to the ease with which they can be completed by patients and results fed back to the treating clinician in very time-limited primary care or low intensity psychological therapy consultations.

Future trials should not limit the measurement of outcome to the data on psychological symptoms, individual functioning and interpersonal relationships provided by the PROMs used as the monitoring intervention (the 'monomethod' as described by Shimokawa 2010). Instead blinded outcome assessors should collect additional data, independently of the treating clinicians, on symptoms and functioning, and also on possible harms, health-related quality of life, social functioning, and costs of the intervention. Studies should characterise their participants systematically, in terms of diagnoses, using standard classifications such as the ICD or DSM criteria.

Post-hoc analyses of subgroups of participants identified early in treatment as either 'on-track' (OT) or 'not on track' (NOT) for a good clinical response suggest that monitoring with the OQ-45 or ORS might improve outcomes for NOT participants, and reduce the number of sessions received by OT participants, but the effect sizes were small, and the quality of evidence for these effects was graded as low. The identified improvement in outcome for NOT participants of a standardised mean difference of -0.22 equates to

a reduction in OQ-45 score of around 4 to 6 points, which is of questionable clinical significance, given the total score on the OQ-45 ranges from 0 to 180, and a difference of 14 points is judged to represent meaningful change (Lambert 2004). No significant differences between feedback and no-feedback groups in the proportions of patients achieving clinically significant change on the OQ-45 or ORS PROMs were reported in most of the studies identified (De Jong 2014; Hawkins 2004; Lambert 2001; Murphy 2012; Reese 2009a; Reese 2009b; Simon 2012), although the numbers of patients achieving significant change in either group were small, and the studies lacked power to determine differences. One identified study (Whipple 2003) did report that significantly more patients in the feedback plus clinical support tool (CST) group achieved clinically significant or reliable change than the no-feedback group. Future studies recruiting larger samples are needed to address the clinical significance of any benefits found.

The identified mean reduction in length of treatment received by OT participants of 0.69 sessions might improve the efficiency of treatment overall, through enabling targeting of therapist time more appropriately to NOT patients, but again the overall difference was small, and none of the studies collected information about costs in relation to the intervention and its effects, so the cost-effectiveness of the approach has not yet been assessed. It should be stressed that these findings among NOT and OT subgroups of participants are the results of post-hoc analyses which were not planned in our original protocol, and should be regarded as hypothesis-forming rather than hypothesis-testing, requiring examination in future, larger, and better-designed studies.

Study designs should be developed which reduce the bias due to patient and clinician awareness that routine outcome monitoring is being applied in the intervention arm, although we acknowledge it is impossible to blind participants and treating clinicians completely since they are being asked to consider the results of PROMs fed back to them. Designs such as Chang 2012 used, which vary the amount and timing of feedback of PROM results, may be helpful in reducing such bias, in addition to utilising a different outcome measure from the PROM being used as the intervention to assess change in symptom scores.

Independent assessment of research outcomes by staff who are not involved in treating the patients should improve outcome assessment, at a specified interval, and reduce the relatively high attrition rates found when follow-up is left entirely to the treating clinician, and no data are collected on patients who do not return. To reduce bias due to incomplete follow-up, researchers should also consider using multiple imputation or other methods to deal with missing data, and report intention to treat in addition to per protocol analyses. Studies are also needed which determine long-term outcomes beyond six months, as only two studies in this review included longer term follow-up.

These measures will make studies more complicated and more costly to conduct, but are needed to address the significant problem of the low quality of evidence overall on the routine outcome monitoring of CMHDs using PROMs.

ACKNOWLEDGEMENTS

Cochrane review group funding acknowledgement:

The National Institute for Health Research (NIHR) is the largest single funder of the Cochrane Common Mental Disorders Group.

Disclaimer:

The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the NIHR, NHS or the Department of Health.

REFERENCES

References to studies included in this review

Amble 2014 *{published data only}*

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. *Psychotherapy Research* 2014;**24**:doi:10.1080/10503307.2014.928756.

Berking 2006 *{published data only}*

* Berking M, Orth U, Lutz W. How effective is systematic feedback of treatment progress to the therapist? An empirical study in a cognitive-behavioural-oriented inpatient setting [Wie effektiv sind systematische Rückmeldungen des Therapieverlaufs an den Therapeuten? Eine empirische studie in einem stationär–verhaltenstherapeutischen setting]. *Zeitschrift Fur Klinische Psychologie Und Psychotherapie* 2006;**35**(1):25–9.

Lutz W, Tholen S, Schürch E, Berking M. [Die entwicklung, validierung und reliabilität von kurzformen gängiger psychometrischer Instrumentezur evaluation des therapeutischenfortschritts in psychotherapieund psychiatrie]. *Diagnostica* 2006;**52**:11–25.

Chang 2012 *{published data only}*

* Chang TE, Jing Y, Yeung AS, Brenneman SK, Kalsekar I, Hebden T, et al. Effect of communicating depression severity on physician prescribing patterns: findings from the Clinical Outcomes in MEasurement-based Treatment

(COMET) trial. *General Hospital Psychiatry* 2012;**34**: 105–12. [DOI: 10.1016/j.genhosppsych.2011.12.003]
Chang TE, Jing Y, Yeung AS, Brenneman SK, Kalsekar ID, Hebden T, et al. Depression monitoring and patient behavior in the Clinical Outcomes in MEasurement-based Treatment (COMET) trial. *Psychiatric Services Hospital & Community Psychiatry* 2014;**65**(8):1058–6.
Yeung AS, Jing Y, Brenneman SK, Chang TE, Baer L, Hebden T, et al. Clinical Outcomes in MEasurement-based Treatment (COMET): A trial of depression monitoring and feedback to primary care physicians. *Depression and Anxiety* 2012;**29**:865–73.

De Jong 2012 ***{published and unpublished data}***
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 controlled trial. *Psychotherapy Research* 2012;**22**(4):464–74.

De Jong 2014 ***{published and unpublished data}***
De Jong K, Timman R, Hakkaart-van Roijen L, Vermeulen P, Kooiman K, Passchier J, et al. The effect of outcome monitoring feedback to clinicians and patients in short and long-term psychotherapy: a randomized controlled trial. *Psychotherapy Research* 2014;**24**(6):629–39.

Hansson 2013 ***{published and unpublished data}***
Hansson H, Rundberg J, Österling A, Öjehagen A, Berglund M. Intervention with feedback using Outcome Questionnaire 45 (OQ-45) in a Swedish psychiatric outpatient population. A randomized controlled trial. *Nordic Journal of Psychiatry* 2013;**67**(4):274–81.

Hawkins 2004 ***{published and unpublished data}***
Hawkins EJ, Lambert MJ, Vermeersch DA, Slade KL, Tuttle KC. The therapeutic effects of providing patient progress information to therapists and patients. *Psychotherapy Research* 2004;**14**(3):308–27.

Lambert 2001 ***{published data only}***
Lambert MJ, Whipple JL, Smart DW, Vermeersch DA, Nielsen SL. The effects of providing therapists with feedback on patient progress during psychotherapy: Are Outcomes Enhanced?. *Psychotherapy Research* 2001;**11**(1):49–68.

Mathias 1994 ***{published data only}***
* Mathias SD, Fifer SK, Mazonson PD, Lubeck DP, Buesching DP, Patrick DL. Necessary but not sufficient: the effect of screening and feedback on outcomes of primary-care patients with untreated anxiety. *Journal of General Internal Medicine* 1994;**9**:606–15.
Mazonson PD, Mathias SD, Fifer SK, Buesching DP, Malek P, Patrick DL. The mental health patient profile: does it change primary-care physicians' practice patterns?. *Journal of the American Board of Family Practice/American Board of Family Practice* 1996;**9**:336–45.

Murphy 2012 ***{published data only}***
Murphy KP, Rashleigh CM, Timulak L. The relationship between progress feedback and therapeutic outcome in student counselling: a randomised control trial. *Counselling Psychology Quarterly* 2012;**25**:1–18.

Probst 2013 ***{published data only}***
Probst T, Lambert MJ, Dahlbender RW, Loew TH, Tritt K. Providing patient progress feedback and clinical support tools to therapists: is the therapeutic process of patients on-track to recovery enhanced in psychosomatic in-patient therapy under the conditions of routine practice?. *Journal of Psychosomatic Research* 2014;**76**(6):477–84.
* Probst T, Lambert MJ, Loew TH, Dahlbender RW, Gollner R, Tritt K. Feedback on patient progress and clinical support tools for therapists: Improved outcome for patients at risk of treatment failure in psychosomatic in-patient therapy under the conditions of routine practice. *Journal of Psychosomatic Research* 2013;**75**(3):255–61.

Reese 2009a ***{published and unpublished data}***
Reese RJ, Norsworthy LA, Rowlands SR. Does a continuous feedback system improve psychotherapy outcome?. *Psychotherapy* 2009;**46**(4):418–31.

Reese 2009b ***{published and unpublished data}***
Reese RJ, Norsworthy LA, Rowlands SR. Does a continuous feedback system improve psychotherapy outcome?. *Psychotherapy* 2009;**46**(4):418–31.

Scheidt 2012 ***{published data only}***
* Scheidt CE, Brockmann J, Caspar F, Rudolf G, Stangier U, Vogel H. The project of the Techniker-Krankenkasse: some comments on the results by the project's scientific advisory board [Das modellprojekt der Techniker–Krankenkasse: eine kommentierung der ergebnisse aus der aicht des wissenschaftlichen projektbeirates]. *Psychotherapie Psychosomatik Medizinische Psychologie* 2012;**62**(11): 405–12.
Wittmann WW, Lutz W, Steffanowski A, Kriz D, Glahn EM, Völkle MC, et al. *Qualitätsmonitoringin der ambulanten Psychotherapie: Modellprojekt der Techniker Krankenkasse - Abschlussbericht*. Hamburg: Techniker Krankenkasse, 2011.

Simon 2012 ***{published data only}***
Simon W, Lambert MJ, Harris MW, Busath G, Vazquez A. Providing patient progress information and clinical support tools to therapists: effects on patients at risk of treatment failure. *Psychotherapy Research* 2012;**22**:638–47.

Trudeau 2001 ***{published and unpublished data}***
Trudeau LS. Effects of a clinical feedback system on client and therapist outcomes in a rural community mental health center. *Dissertation Abstracts International: Section B: The Sciences and Engineering Dissertation Abstracts International* 2001;**61**:5232.

Whipple 2003 ***{published data only}***
Whipple JL. Improving the effects of psychotherapy: the use of early identification of treatment failure and problem solving strategies in outcome [thesis]. *Dissertation Abstracts International: Section B: Sciences and Engineering* 2002;**63** (6-B):3030.
* Whipple JL, Lambert MJ, Vermeersch DA, Smart DW, Nielsen SL, Hawkins EJ. Improving the effects of psychotherapy: the use of early identification of treatment

failure and problem-solving strategies in routine practice. *Journal of Counseling Psychology* 2003;**50**(1):59–68.

References to studies excluded from this review

Anker 2009 *{published data only}*
Anker MG, Duncan BL, Sparks JA. Using client deedback to improve couple therapy outcomes: a randomized clinical trial in a naturalistic setting. *Journal of Consulting and Clinical Psychology* 2009;**77**(4):693–704.

Brodey 2005 *{published data only}*
Brodey BB, Cuffel B, McCulloch J, Tani S, Maruish M, Brodey I, et al. The acceptability and effectiveness of patient-reported assessments and feedback in a managed behavioral healthcare setting. *The American Journal of Managed Care* 2005;**11**(12):774–80.

Brody 1990 *{published data only}*
Brody DS, Lerman CE, Wolfson HG, Caputo GC. Improvement in physicians' counseling of patients with mental health problems. *Archives of Internal Medicine* 1990; **150**(5):993–8.

Dobscha 2006 *{published data only}*
Dobscha SK, Corson K, Hickam DH, Perrin NA, Kraemer DF, Gerrity MS. Depression decision support in primary care: a cluster randomized trial. *Annals of Internal Medicine* 2006;**145**:477–87.

Fluckiger 2012 *{published data only}*
Fluckiger C, Del Re AC, Wampold BE, Znoj H, Caspar F, Jorg U. Valuing clients' perspective and the effects on the therapeutic alliance: a randomized controlled study of an adjunctive instruction. *Journal of Counseling Psychology* 2012;**59**(1):18–26.

Haderlie 2012 *{published data only}*
Haderlie MM. Enhancing therapists' clinical judgments of client progress subsequent to objective feedback. *Dissertation abstracts International: Section B: The Sciences and Engineering Dissertation Abstracts International* 2012;**73**: 616.

Harmon 2007 *{published data only}*
Harmon SC, Lambert MJ, Smart DM, Hawkins E, Nielsen SL, Slade K, et al. Enhancing outcome for potential treatment failures: therapist client feedback and clinical support tools. *Psychotherapy Research* 2007;**17**(4):379–92.

Lambert 2002 *{published data only}*
Lambert MJ, Whipple JL, Vermeersch DA, Smart DW, Hawkins EJ, Nielsen SL, et al. Enhancing psychotherapy outcomes via providing feedback on client progress: a replication. *Clinical Psychology & Psychotherapy* 2002;**9**(2): 91–103.

Newnham 2010 *{published data only}*
Newnham EA, Hooke GR, Page AC. Progress monitoring and feedback in psychiatric care reduces depressive symptoms. *Journal of Affective Disorders* 2010;**127**:139–46.

Pedersen 2014 *{published data only}*
Pedersen ER, Kaysen DL, Lindgren KP, Blayney J, Simpson TL. Impact of daily assessments on distress and PTSD symptoms in trauma-exposed women. *Journal of Interpersonal Violence* 2014;**29**(5):824–45.

Priebe 2007 *{published data only}*
Priebe S, McCabe R, Bullenkamo J, Hansson L, Lauber C, Martinez-Leal R, et al. Structured patient-clinician communication and 1-year outcome in community mental healthcare. *British Journal of Psychiatry* 2007;**191**(5):420–6.

Puschner 2009 *{published data only}*
Puschner B, Schofer D, Knaup C, Becker T. Outcome management in in-patient psychiatric care. *Acta Psychiatrica Scandinavica* 2009;**120**:308–19.

Reese 2010 *{published data only}*
Reese RJ, Toland MD, Slone NC, Norsworthy LA. Effect of client feedback on couple psychotherapy outcomes. *Psychotherapy* 2010;**47**(4):616–30.

Reese 2013 *{published data only}*
Reese RJ, Gillaspy JA, Owen JJ, Flora KL, Cunningham LC, Archie D, et al. The influence of demand characteristics and social desirability on clients' ratings of the therapeutic alliance. *Journal of Clinical Psychology* 2013;**69**(7):696–709.

Reeves 2010 *{published data only}*
Reeves EK. The use of psychotherapy feedback on client progress: implications for training and clinical practice. *The use of psychotherapy feedback on client progress: implications for training and clinical practice. PhD Thesis, New York University*. Ann Arbor, MI: Proquest LLC, 2010.

Rise 2012 *{published data only}*
Rise MB, Eriksen L, Grimstad H, Steinsbekk A. The short-term effect on alliance and satisfaction of using patient feedback scales in mental health out-patient treatment. A randomised controlled trial. *BMC Health Services Research* 2012;**12**:348.

Simon 2000 *{published data only}*
Simon GE, VonKorff M, Rutter C, Wagner E. Randomised trial of monitoring, feedback, and management of care by telephone to improve treatment of depression in primary care. *BMJ (Clinical Research ed)* 2000;**320**:550–4.

Slade 2006 *{published data only}*
Slade M, McCrone P, Kuipers E, Leese M, Parabiaghi A, Priebe S, et al. Use of standardised outcome measures in adult mental health services:randomised controlled trial. *British Journal of Psychiatry* 2006;**189**:330–6.

Slade 2008 *{published data only}*
Slade K, Lambert MJ, Harmon SC, Smart DW, Bailey R. Improving psychotherapy outcome: the use of immediate electronic feedback and revised clinical support tools. *Clinical Psychology & Psychotherapy* 2008;**15**(5):287–303.

References to studies awaiting assessment

Gibbons 2015 *{published data only}*
* Gibbons M, Kurtz J, Thompson D, Mack R, Lee J, Rothbard A, et al. The effectiveness of clinician feedback in the treatment of depression in the community mental

health system. *Journal of Consulting and Clinical Psychology.* 2015;**83**:748–59.

Gibbons MB. The development of a therapist feedback system for MDD in community mental health [protocol]. *NIH Research Portfolio Online Reporting Tools (RePORT) [projectreporter.nih.gov/project˙info˙description.cfm?aid=7788979]* [Accessed 26 May 2016] 2010.

Rise 2016 ***{published data only}***

Rise MB, Eriksen L, Grimstada H, Steinsbekk A. The long-term effect on mental health symptoms and patient activation of using patient feedback scales in mental health out-patient treatment. A randomised controlled trial. *Patient Education and Counselling* 2016;**99**:164–8.

References to ongoing studies

Metz 2015 ***{published data only}***

Metz MJ, Franx GC, Veerbeek MA, De Beurs E, Van der Feltz-Cornelis CM, Beekman AT. Shared decision making in mental health care using routine outcome monitoring as a source of information: a cluster randomised controlled trial. *BMC Psychiatry* 2015;**15**:313.

NCT01796223 ***{published data only}***

Linaker OM, Lara MM. Effects of systematic patient feedback on therapy outcome and dropout: a randomized controlled study on adult out-patients at a community mental health centre. clinicaltrials.gov/ct2/show/NCT01796223 2013 [Accessed 24 November 2015].

NCT02023736 ***{published data only}***

Pinsof W. Assessing psychotherapy outcome in treatment as usual versus treatment as usual with the STIC feedback system. clinicaltrials.gov/ct2/show/NCT02023736 2013 [Accessed 24 November 2015].

NCT02095457 ***{published data only}***

Biran L, Ganor O. A randomized trial of routine computerized outcome and process clinical measures monitoring in mental health outpatient services: preparing for the planned public mental health reform in Israel. clinicaltrials.gov/ct2/show/NCT02095457 2014 [Accessed 24 November 2015].

NCT02656641 ***{published data only}***

Ludwig KG. Using the Patient Health Questionnaire-9 (PHQ-9) and Generalized Anxiety Disorder 7-Item Scale (GAD-7) as Feedback Instruments in Brief Psychotherapy. *clinicaltrials.gov/ct2/show/NCT02656641 [Accessed 26 May 2016]* 2015.

NTR5466 ***{published data only}***

Bovendeerd AM. Routine Process Monitoring, systematic patient feedback in the primary and specialised mental healthcare - RPM. *trialregister.nl/trialreg/admin/rctview.asp?TC=5466 [Accessed 26 May 2016]* 2015.

NTR5707 ***{published data only}***

Bastiaansen JACJ. Self-monitoring and personalized feedback as a tool to boost depression treatment. *trialregister.nl/trialreg/admin/rctview.asp?TC=5707 [Accessed 26 May 2016]* 2016.

Additional references

Andlin-Sobocki 2005

Andlin-Sobocki P, Jonsson B, Wittchen HU, Olesen J. Cost of disorders of the brain in Europe. *European Journal of Neurology* 2005;**12**:1–27.

APA 2000

American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision (DSM-IV-TR)*. Washington, DC: American Psychiatric Association, 2000.

Archer 2012

Archer J, Bower P, Gilbody S, Lovell K, Richards D, Gask L, et al. Collaborative care for depression and anxiety problems. *Cochrane Database of Systematic Reviews* 2012, Issue 10. [DOI: 10.1002/14651858.CD006525.pub2]

Barkham 2006

Barkham M, Mellor-Clark J, Connell J, Cahill, J. A core approach to practice-based evidence: a brief history of the origins and applications of the CORE-OM and CORE System. *Counselling & Psychotherapy Research* 2006;**6**(1): 3–15.

Beck 1961

Beck AT, Ward CH. An inventory for measuring depression. *Archives of General Psychiatry* 1961;**4**:561–71.

Beck 1996

Beck AT, Steer RA, Brown GK. *Manual for the Beck Depression Inventory-II*. San Antonio, TX: Psychological Corporation, 1996.

Black 2013

Black N. Patient reported outcome measures could help transform healthcare. *BMJ* 2013;**346**:doi:10.1136/bmj.f167.

Black 2015

Black N, Burke L, Forrest CB, Ravens Sieberer UH, Ahmed S, Valderas JM, et al. Patient-reported outcomes: pathways to better health, better services, and better societies. *Quality of Life Research* 2015;**24**:doi:10.1007/s11136–015-1168-3.

Bland 1997

Bland JM, Kerry SM. Statistics notes. Trials randomised in clusters. *BMJ* 1997;**710**:600.

BMA & NHS Employers 2009

British Medical Association & NHS Employers. *Quality and Outcomes Framework guidance for GMS contract 2009/10. Delivering investment in general practice*. London: NHS Confederation, 2009.

Bostwick 2000

Bostwick JM, Pankratz VS. Affective disorders and suicide risk: a reexamination. *American Journal of Psychiatry* 2000; **157**:1925–32.

Boyce 2013

Boyce MB, Browne JP. Does providing feedback on patient-reported outcomes to healthcare professionals result in better outcomes for patients? A systematic review. *Quality of Life Research* 2013;**22**(9):2265-78 doi:10.1007/s11136-013-0390-0.

Carlier 2012
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. *Journal of Evaluation in Clinical Practice* 2012;**18** (1):104–10.

Compton 2006
Compton WM, Conway KP, Stinson FS, Grant BF. Changes in the prevalence of major depression and comorbid substance use disorders in the United States between 1991-1992 and 2001-2002. *American Journal of Psychiatry* 2006; **163**:2141–7.

Cooper 1982
Cooper P, Osborne M, Gath D, Feggetter G. Evaluation of a modified self-report measure of social adjustment. *British Journal of Psychiatry* 1982;**141**:68–75.

Crits-Christoph 2012
Crits-Christoph P, Ring-Kurtz S, Hamilton JL, Lambert MJ, Gallop R, McClure B, et al. A preliminary study of the effects of individual patient-level feedback in outpatient substance abuse treatment programs. *Journal of Substance Abuse Treatment* 2012;**42**:301–9.

Davidson 2014
Davidson K, Perry A, Bell L. 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* 2014;**88**(1): 21–37.

Deeks 2011
Deeks JJ, Higgins JPT, Altman DG (editors). Chapter 9: Analysing data and undertaking meta-analyses. In: Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 (updated March 2011). The Cochrane Collaboration, 2011. Available from cochrane-handbook.org.

Derogatis 1974
Derogatis LR, Lipman RS, Rickels K, Uhlenhuth EH, Covi L. The Hopkins Symptom Checklist (HSCL): a self-report symptom inventory. *Behavioural Sciences* 1974;**19**(1):1–15.

Derogatis 1983
Derogatis LR. *SCL-90-R: Administration, Scoring and Procedures Manual-II*. Baltimore, Md: Clinical Psychometric Research, 1983.

Divine 1992
Divine GW, Brown JT, Frazer LM. The unit of analysis error in studies about physicians' patient care behavior. *Journal of General Internal Medicine* 1992;**7**:623–9.

Dolan 1997
Dolan P. Modeling valuations for EuroQol health states. *Medical Care* 1997;**35**(11):1095–108.

Dowrick 2009
Dowrick C, Leydon GM, McBride A, Howe A, Burgess H, Clarke P, et al. Patients' and doctors' views on depression severity questionnaires incentivised in UK quality and outcomes framework: qualitative study. *BMJ* 2009;**338**: b663. doi:10.1136/bmj.b663.

Egger 1997
Egger M, Davey-Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple graphical test. *BMJ* 1997;**315**:629–35.

Ferrari 2013
Ferrari AJ, Charlson FJ, Norman RE, Patten SB, Freedman G, Murray CJL, et al. Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010. *PLoS Medicine* 2013;**10**(11): e1001547.doi:10.1371/journal.pmed.1001547.

First 1997
First M, Spitzer R, Gibbon M, Williams J. Structured clinical interview for DSM-IV axis I disorders (SCIDI), clinician version. *Structured clinical interview for DSM-IV axis I disorders (SCIDI), clinician version*. Washington: American Psychiatric Press, 1997.

Fitzpatrick 2009
Fitzpatrick R, Gibbons E, Mackintosh A. *An overview of patient-reported outcome measures for people with anxiety and depression*. Oxford: Patient-reported Outcome Measurement Group, Department of Public Health, University of Oxford, 2009.

Gilbody 2002
Gilbody SM, House AO, Sheldon T. Routine administration of Health Related Quality of Life (HRQoL) and needs assessment instruments to improve psychological outcome - a systematic review. *Psychological Medicine* 2002;**32**(1345): 1356.

Gilbody 2008
Gilbody S, Sheldon T, House A. Screening and case-finding instruments for depression: a meta-analysis. *Canadian Medical Association Journal* 2008;**178**(8):997–1003.

Gilchrist 2007
Gilchrist G, Gunn J. Observational studies of depression in primary care: what do we know?. *BMC Family Practice* 2007;**8**:DOI: 10.1186/1471-2296-8-28.

Goldberg 1972
Goldberg D. The Detection of Psychiatric Illness by Questionnaire. *The Detection of Psychiatric Illness by Questionnaire*. Oxford: Oxford University Press, 1972.

Greenhalgh 2005
Greenhalgh J, Long AF, Flynn R. The use of patient reported outcome measures in routine clinical practice: Lack of impact or lack of theory?. *Social Science and Medicine* 2005; **60**(4):833-43. doi:10.1016/j.socscimed.2004.06.022.

Gulliford 1999
Gulliford MC, Ukoumunne OC, Chinn S. Components of variance and intraclass correlations for the design of community-based surveys and intervention studies. *American Journal of Epidemiology* 1999;**149**:876–83.

Hamilton 1960
Hamilton M. A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry* 1960;**23**:56–62.

Higgins 2003
Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003;**327**: 557–60.

Higgins 2011
Higgins JPT, Altman DG, Sterne JAC (editors). Chapter 8: Assessing risk of bias in included studies. In: Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 (updated March 2011). The Cochrane Collaboration, 2011. Available from cochrane-handbook.org.

HRSA 2005
Health Resources and Services Administration (HRSA) Health Disparities Collaborative on Depression. *Standards on the assessment of patients at diagnosis and follow-up*. qualitymeasures.ahrq.gov/browse/by-organization-indiv.aspx?objid=26124(accessed 18 November 2013): AHRQ, 2005.

HSCIC 2015
Community and Mental Health Team. *Psychological Therapies: Annual Report on the use of IAPT services England, 2014/15*. gov.uk/government/uploads/system/uploads/attachment˙data/file/386221/NHS˙England˙Mandate.pdf: Health and Social Care Information Centre, [Accessed 26 November 2015] 2015.

IAPT 2011
The IAPT Data Handbook NHS. *Guidance on recording and monitoring outcomes to support local evidence-based practice*. Accessed 18 November 2013: iapt.nhs.uk/silo/files/the-iapt-data-handbook.pdf NHS, [Accessed 1 May 2014] 2011.

Katon 1992
Katon W, Schulberg H. Epidemiology of depression in primary care. *General Hospital Psychiatry* 1992;**14**(4): 237–47.

Kendrick 2009
Kendrick T, Dowrick C, McBride A, Howe A, Clarke P, Maisey S, et al. Management of depression in UK general practice in relation to scores on depression severity questionnaires: analysis of medical record data. *BMJ* 2009; **338**:b750. doi:10.1136/bmj.b750.

Kendrick 2014
Kendrick T, Moore M, Gilbody S, Churchill R, Stuart B, El-Gohary M. Routine use of patient reported outcome measures (PROMs) for improving treatment of common mental health disorders in adults. *Cochrane Database of Systematic Reviews* 2014, Issue 6. [DOI: 10.1002/14651858.CD011119]

Kessler 2005
Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey replication. *Archives of General Psychiatry* 2005;**62**:593–602.

King 2008
King M, Nazareth I, Levy G, Walker C, Morris R, Weich S, et al. Prevalence of common mental disorders in general practice attendees across Europe. *British Journal of Psychiatry* 2008;**192**:362–7.

Knaup 2009
Knaup C, Koesters M, Schoefer D, Becker T, Puschner B. Effect of feedback of treatment outcome in specialist mental healthcare: meta-analysis. *British Journal of Psychiatry* 2009; **195**:15-22 doi:10.1192/bjp.bp.108.053967.

Krageloh 2015
Krageloh CU, Czuba KJ, Billington DR, Kersten P, Siegert RJ. Using feedback from patient-reported outcome measures in mental health services: a scoping study and typology. *Psychiatric Services* 2015;**66**(3):224–41.

Lambert 2003
Lambert MJ, Whipple JL, Hawkins EJ, Vermeersch DA, Nielsen SL, Smart DW. Is it time for clinicians to routinely track patient outcome? A meta-analysis. *Clinical Psychology: Science and Practice* 2003;**10**:288–301.

Lambert 2004
Lambert MJ, Gregersen AT, Burlingame GM. The Outcome Questionnaire-45. In: Murish ME editor(s). *Use of Psychological Testing for Treatment Planning and Outcome Assessment*. Vol. **3**, New Jersey: Erlbaum, 2004:191–234.

Lewis 1992
Lewis G, Pelosi AJ, Araya R, Dunn G. Measuring psychiatric disorder in the community: a standardized assessment for use by lay interviewers. *Psychological Medicine* 1992;**22**(2): 465–86.

Malpass 2010
Malpass A, Shaw A, Kessler D, Sharp D. Concordance between PHQ-9 scores and patients' experiences of depression: a mixed methods study. *British Journal of General Practice* 2010;**60**:doi:10.3399/ bjgp10X502119.

Marciniak 2004
Marciniak M, Lage MJ, Landbloom RP, Dunayevich E, Bowman L. Medical and productivity costs of anxiety disorders: case control study. *Depression and Anxiety* 2004; **19**:112–20.

Marshall 2006
Marshall S, Haywood K, Fitzpatrick R. Impact of patient-reported outcome measures on routine practice: A structured review. *Journal of Evaluation in Clinical Practice* 2006;**12**(5):559-68. doi:10.1111/j.1365-2753.2006.00650.x.

McCrone 2008
McCrone P, Dhanasiri S, Patel A, Knapp M. *Paying the price: the cost of mental health care in England to 2026*. London: King's Fund, 2008.

McManus 2009
McManus S, Meltzer H, Brugha T, Bebbington P, Jenkins R. Adult psychiatric morbidity in England, 2007. Results of a household survey. *Adult psychiatric morbidity in England, 2007. Results of a household survey*. Leicester: National

Centre for Social Research & Department of Health Sciences, University of Leicester, 2009.

Miller 2003
Miller SD, Duncan BL, Brown J, Sparks J, Claud D. The Outcome Rating Scale: a preliminary study of the reliability, validity, and feasibility of a brief visual analog measure. *Journal of Brief Therapy* 2003;**2**(2):91–100.

Moher 2009
Moher D, Liberati A, Tetzlaff J, Altman DG, the PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Medicine* 2009;**6**(7):e1000097.

Montgomery 1979
Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. *British Journal of Psychiatry* 1979;**134**(4):382–9.

Moore 2012
Moore M, Ali S, Stuart B, Leydon GM, Ovens J, Goodall C, et al. Depression management in primary care: an observational study of management changes related to PHQ-9 score for depression monitoring. *British Journal of General Practice* 2012;**62**:310–1.

Mundt 2002
Mundt JC, Marks IM, Shear MK, Greist JM. The Work and Social Adjustment Scale: a simple measure of impairment in functioning. *British Journal of Psychiatry* 2002;**180**:461-4. doi: 10.1192/bjp.180.5.461.

Murray 2010
Murray CJL, Vos T, Lozano R, Naghavi M, Flaxman AD, Michaud C, et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* 2012;**380**(9859):2197–223.

NICE 2009
National Institute for Health and Clinical Excellence. National Collaborating Centre for Mental Health. *Depression. The treatment and management of depression in adults. NICE Clinical Guideline 90*. Updated from 2004 guideline. London: NICE, 2009.

NICE 2011a
National Institute for Health and Clinical Excellence. *NICE Guideline: common mental health disorders: identification and pathways to care; CG123*. London: British Psychological Society and Royal College of Psychiatrists, 2011.

NICE 2011b
National Institute for Health & Clinical Excellence. *Quality Standards for the Management of Depression in Adults. publications.nice.org.uk/depression-in-adults-quality-standard-qs8/quality-statement-3-recording-health-outcomes (Accessed 18 November 2013)*. London: NICE, 2011.

Poston 2010
Poston JM, Hanson WE. Meta-analysis of psychological assessment as a therapeutic intervention. *Psychological Assessment* 2010;**22**(2):203–12.

QOF Database 2013
QOF Database. UK level data for Quality and Outcomes Framework 2008-2013 depression indicators. gpcontract.co.uk/browse/UK/Depression/13 (accessed 21 November 2013).

Radloff 1997
Radloff LS. The CES-D scale a self-report depression scale for research in the general population. *Applied Psychological Measurement* 1977;**1**:385–401.

RevMan 2014 [Computer program]
The Nordic Cochrane Centre, The Cochrane Collaboration. Review Manager (RevMan). Version 5.3. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2014.

Robins 1981
Robins NL, Helzer JE, Croughan J, Ratcliff KS. The National Institute of Mental Health Diagnostic Interview Schedule: its History, Characteristics, and Validity. *Archives of General Psychiatry* 1981;**38**:381–99.

Shaw 2013
Shaw EJ, Sutcliffe D, Lacey T, Stokes T. Assessing depression severity using the UK Quality and Outcomes Framework depression indicators: a systematic review. *British Journal of General Practice* 2013;**63**:doi:10.3399/bjgp13X667169.

Sheehan 1998
Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.) the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry* 1998;**Suppl 20**:22–3.

Shimokawa 2010
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. *Journal of Consulting and Clinical Psychology* 2010;**78**:298–311.

Simon 1997
Simon GE, Katzelnick DJ. Depression, use of medical services and cost-offset effects. *Journal of Psychosomatic Research* 1997;**42**:333–4.

Simon 2004
Simon G, Ludman E, Tutty S, Operskalski B, Von Korff M. Telephone psychotherapy and telephone care management for primary care patients starting antidepressant treatment:a randomized controlled trial. *JAMA* 2004;**292**:935–42.

Simon 2013
Simon W, Lambert MJ, Busath G, Vazquez A, Berkeljon A, Hyer K, et al. Effects of providing patient progress feedback and clinical support tools to psychotherapists in an inpatient eating disorders treatment program: a randomized controlled study. *Psychotherapy Research* 2013;**23**:287–300.

Spitzer 1999
Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of the PRIME-MD: the PHQ primary care study. *JAMA* 1999;**282**(18):1737–44.

Spitzer 2006
Spitzer RL, Kroenke K, Williams JBW, et al. A brief measure for assessing generalised anxiety disorder: the GAD-7. *Archives of Internal Medicine* 2006;**166**:1092–7.

StataCorp. 2015 [Computer program]
College Station, TX: StataCorp LP. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP, 2015.

Trivedi 2004
Trivedi MH, Rush AJ, Ibrahim HM, Carmody TJ, Biggs MM, Suppes T, et al. The Inventory of Depressive Symptomatology, Clinician Rating (IDS-C) and the Quick Inventory of Depressive Symptomatology,Clinician Rating (QIDS-C) and Self Report (QIDS-SR) in public sector patients with mood disorders: a psychometric evaluation. *Psychological Medicine* 2004;**34**:73–82.

Valdera 2008
Valderas JM, Kotzeva A, Espallargues M, Guyatt G, Ferrans CE, Halyard MY, et al. The impact of measuring patient-reported outcomes in clinical practice: a systematic review of the literature. *Quality of Life Research* 2008;**17**:179-93 doi:10.1007/s11136-007-9295-0.

Ware 1996
Ware JE, Kosinski M, Keller SD. A 12-item short-form health survey construction of scales and preliminary tests of reliability and validity. *Medical Care* 1996;**34**:220–33.

Wells 1989
Wells KB, Stewart A, Hays RD, Burnam MA, Rogers W, Daniels M, et al. The functioning and well-being of depressed patients. Results from the Medical Outcomes Study. *JAMA* 1989;**262**:914–9.

Wetherall 2005
Wetherell JL, Gatz M. The Beck Anxiety Inventory in older adults with generalized anxiety disorder. *Journal of Psychopathology and Behavioral Assessment* 2005;**27**(1): 17–24.

WHO 1992
World Health Organization (WHO). *The ICD-10 Classification of Mental and Behavioural Disorders*. Geneva: WHO, 1992.

Yeung 2012
Yeung AS, Jing Y, Brenneman SK, Chang TE, Baer L, Hebden T, et al. Clinical Outcomes in MEasurement-based Treatment (COMET): a trial of depression monitoring and feedback to primary care physicians. *Depression and Anxiety* 2012;**29**(10):865–73.

Zigmond 1983
Zigmond AS, Snaith RP. The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica* 1983;**67**(6): 361–70.

Zung 1965
Zung WW. A self rating depression scale. *Archives of General Psychiatry* 1965;**12**:63–70.

* *Indicates the major publication for the study*

CHARACTERISTICS OF STUDIES

Characteristics of included studies *[ordered by study ID]*

Amble 2014

Methods	**Study design:** Individual randomised controlled trial **Setting:** 2 inpatient clinics and 4 outpatient clinics in mental health care institutions **Country:** Norway
Participants	**Diagnosis:** • 47% Affective disorder • 33% Anxiety disorder • 7% ADHD • 4% Substance abuse • 4% Eating disorders • 3% Personality disorders • 1% Schizophrenia • 1% No diagnosis **Method of diagnosis:** Routine diagnosis by the treating therapist, using the ICD-10 **Age:** Mean age 35.8, SD 11.6, range 18-65 **Sex:** 231 (68%) female **Number:** 377 invited, 340 accepted, 321 randomised, (feedback group 174, controls 147), of whom 259 (81%) were followed up (feedback group 144, controls 115) **Inclusion criteria:** • Attending IP or OP psychiatric clinic for a minimum of two sessions and willing to complete outcome measures. **Exclusion criteria:** • Fewer than 2 outcome questionnaire administrations • Inability to complete the OQ-45 **Co-morbidities:** • Not stated **Losses to follow-up/withdrawal:** 19 excluded as incorrectly randomised; 62 (19%) failed to complete follow-up outcome measures: 14 in the feedback group failed to complete the initial OQ-45, 15 in the non-feedback group 16 in the feedback group only completed one OQ-45 questionnaire, 17 in the non-feedback group **Demographics considered:** Not stated, beyond age and gender **Ethnicity:** Not stated
Interventions	**PROM used as intervention**: Outcome Questionnaire 45 (OQ-45) Participants were randomly assigned to either: **1) Feedback to therapist** **Duration:** variable, number of clinic visits determined by therapist Therapist given feedback prior to seeing participant. (Therapist free to discuss feedback with participant) **2) Control group** **Duration:** variable, number of clinic visits determined by therapist The comparison group also completed OQ-45s but their scores were kept hidden from

	the therapists and participants
Outcomes	**Time points for assessment:** at last clinic visit **Outcomes of the trial** (as reported): • Change in OQ-45 total score* • Proportions recovered, improved, unchanged, and deteriorated • Effect of clinic type on outcome **Subgroups:** Effect of being a 'signal' case (not on track) examined in general linear modelling *outcomes prespecified for this review
Notes	**Dates of study:** Inclusion period June 2010-September 2013 **Funding:** Not stated

Risk of bias

Bias	**Authors' judgement**	**Support for judgement**
Random sequence generation (selection bias)	Low risk	"patients were randomized into the FB or NFB conditions in blocks of 8 and by gender" (p3)
Allocation concealment (selection bias)	Unclear risk	No details provided
Blinding of participants and personnel (performance bias) All outcomes	High risk	"The OQ-Analyst software provides the therapist and patient with a report showing the session-by-session progress" (p3)
Blinding of outcome assessment (detection bias) All outcomes	High risk	The PROM used for feedback was also used for outcome assessment, so the participants themselves were the outcome assessors and they were not blind to whether or not they received the intervention
Incomplete outcome data (attrition bias) All outcomes	High risk	19% failed to complete the outcome measure, with an imbalance between groups (30/174 (17%) in feedback group versus 32/147 (22%) in the non-feedback group
Selective reporting (reporting bias)	Low risk	No evidence of selective reporting found
Other bias	Unclear risk	Only 25% of the therapists employed at the clinics agreed to participate, and three quarters of the patients included were seen in the clinic where the project leader and main coordinator worked, suggesting that the therapists seeing clients in that study were self-selected for their enthusiasm for

		routine outcome monitoring

Berking 2006

Methods	**Study design:** Individual randomised controlled trial **Setting:** Inpatient psychotherapy **Country:** Germany
Participants	**Diagnosis:** • 33% Depressive disorders • 23% Anxiety disorders • 19% Adjustment • 25% not recorded Mean (SD) number of F-diagnoses according to ICD-10: Intervention group 1.59 (0. 90), Control group 1.66 (0.89) **Method of diagnosis:** Clinician diagnosis according to ICD-10 **Age:** mean 49.41 years, SD 8.63. Range not reported **Sex:** 73 (61.9%) female **Number:** 118 randomised, (58 intervention group, 60 control group) **Inclusion criteria:** Consecutive admissions to inpatient psychotherapy **Exclusion criteria:** Not reported **Co-morbidities:** Not reported **Losses to follow-up/withdrawals:** None (all inpatients, final assessment conducted at discharge) • Pre data sets: Intervention group: 88%, Control group: 78%. Pre data sets for CGI, VEV: Intervention group: 98%, Control group: 100% • Post data sets: Intervention group: 81%, Control group: 77% Post data sets for CGI, VEV: Intervention group: 97%, Control group: 98% **Demographics considered:** Age and gender only **Ethnicity:** Not reported
Interventions	**PROM used as intervention**: Questionnaire for assessing success and course of psychotherapeutic treatment (FEV) Participants were randomly assigned to either: **Intervention group (58):** Mean values of FEV at admission were converted into T-scores and were presented together with the percentage of goal attainment (agreed on at admission) on a feedback form Feedback was provided to therapists the following working day. Feedback to clinician only, but they were allowed to discuss results with participants **Control group (60):** No feedback to clinician or participant
Outcomes	**Outcomes:** • Set of short forms for assessing success and course of psychotherapy* FEV and FEP: Fragebögen zur Erfassung von Erfolgen und Verläufen psychotherapeutischer Behandlungen (Lutz et al., 2006) including: Short form of Emotionalitätsinventar (EMI-B); Short form of Brief Symptom Inventory (BSI); Short form of Inventar zur Erfassung Interpersonaler Probleme (IIP-D); and Short form of Inkongruenzfragebogens (INK)

	• Clinical Global Impression (CGI) • Changes in experience and behaviour (VEV: Veränderungen des Erlebens und Verhaltens) **Time points for assessment:** FEV and FEP values were assessed at admission, 2 days later and on a weekly basis. CGI and VEV were assessed at discharge *outcomes prespecified for this review
Notes	**Dates of study:** Not stated **Funding:** Not stated

Risk of bias

Bias	**Authors' judgement**	**Support for judgement**
Random sequence generation (selection bias)	High risk	Quote: "Der Randomisierungsprozess erfasste alle Patienten und erfolgte per Münzwurf" (tossing a coin) p.23
Allocation concealment (selection bias)	High risk	Quote: "Der Randomisierungsprozess erfasste alle Patienten und erfolgte per Münzwurf" (tossing a coin) p.23
Blinding of participants and personnel (performance bias) All outcomes	High risk	Not possible to blind participants and personnel due to the nature of the intervention
Blinding of outcome assessment (detection bias) All outcomes	Unclear risk	Not reported
Incomplete outcome data (attrition bias) All outcomes	Low risk	• Pre data sets: Intervention group: 88%, Control group: 78%. Pre data sets for CGI, VEV: Intervention group: 98%, Control group: 100% • Post data sets: Intervention group: 81%, Control group: 77% Post data sets for CGI, VEV: Intervention group: 97%, Control group: 98%
Selective reporting (reporting bias)	Low risk	No evidence of selective reporting found
Other bias	Low risk	No other perceived bias

Chang 2012

Methods	**Study design:** Cluster randomised controlled trial **Setting:** Primary care across 74 sites **Country:** USA
Participants	**Diagnosis:** Major depressive disorder **Method of diagnosis:** Not specified **Age:** Intervention group, M = 46.6 (SD = 15.0); control group, M = 45.3 (SD = 15.4); range = 18-65+ **Sex:** 216 Male, 425 Female, 274 not reported **Number:** 915 randomised, 642 in analysis (364 intervention group, 278 control group) **Inclusion criteria:** • Physician diagnosis of major depressive disorder • Being capable of self-management • Sufficient comprehension of English to complete surveys and telephone interviews **Exclusion criteria:** • Antidepressant use within previous 120 days • Baseline PHQ score < 5 • Bereavement < 8 weeks prior to enrolment • Current postpartum depression or pregnancy • Need for psychiatric hospitalisation at enrolment visit • History of psychotic disorder • History of bipolar disorder • History of suicide attempts or current suicide plan • Previous electroconvulsive therapy • Previous vagus nerve stimulation • Previous transcranial magnetic stimulation • Previous magnetic seizure therapy • Previous deep brain stimulation **Co-morbidities:** • Anxiety disorder • Chronic pain • 'Other co-morbidity' **Losses to follow-up/withdrawals:** • Intervention group 139/503 (27.6%). Did not complete follow up surveys = 42, did not participate in 6 month interview = 81. Baseline PHQ < 5 = 16 • Control group 134/412 (32.5%). Did not complete baseline survey = 1, did not complete follow up surveys = 85, did not participate in 6 month interview = 42. Baseline PHQ < 5 = 6 **Demographics considered:** • Region • Urbanicity • Type of insurance • Employment status • Education • Marital status **Ethnicity:** • White 484 (75.4%) • Black 110 (17.1%) • Asian 7 (1.1%)

	• Native Hawaiian or Pacific Islander 2 (0.3%) • American Indian or Alaska native 8 (1.2%) • Hispanic/Latino 74 (11.6%) • Other 32 (5.0%) (Patients could indicate multiple options for ethnicity; groups are not mutually exclusive)
Interventions	**PROM used as intervention**: PHQ-9 Participants were randomly assigned to either: **Intervention group (503)** **Duration:** 6 months, number of clinic visits determined by physician PHQ-9 scores of each participant faxed to physicians on a monthly basis along with previous scores obtained, percentage change in baseline score, criteria for interpreting the results, general reminders and possible treatment adjustments **Control group (412)** **Duration:** 6 months, number of clinic visits determined by physician PHQ-9 scores of each participant faxed to physicians after 6 months (end of study period) along with previous scores obtained, percentage change in baseline score, criteria for interpreting the results, general reminders and possible treatment adjustments
Outcomes	**Time points for assessment:** (7), Baseline, months 1-6 **Outcomes of the trial** (as reported): • Remission (PHQ score < 5)* • Response (PHQ score reduced by at least 25%)* • Pharmacological treatment patterns* • Reports of self harm/suicide* • Physician use of interview results *outcomes prespecified for this review
Notes	**Dates of study:** 2009-2010 **Sources of funding:** Bristol-Myers Squibb, Otsuka Pharmaceutical Co., Ltd Characteristics and data obtained from both Chang 2012 (primary reference) and Yeung 2012

Risk of bias

Bias	Authors' judgement	Support for judgement
Random sequence generation (selection bias)	High risk	"Investigator sites were alternately (1:1) cluster-assigned to usual care and intervention arms prior to patient enrolment." (p. 106) We judged this at a high risk of bias due to the alternate assignment of sites
Allocation concealment (selection bias)	High risk	See comments above in 'Random sequence generation'

Chang 2012 (*Continued*)

Blinding of participants and personnel (performance bias) All outcomes	Low risk	"Physicians were blinded as to which study arm their practice was assigned to, and all physicians were not informed of the frequency at which patient status reports would be delivered for either arm." (p. 867 Yeung 2012). We judged this as adequate blinding given the study design
Blinding of outcome assessment (detection bias) All outcomes	Low risk	Outcome was measured using the PROM used for feedback, but "Physicians were blinded as to which study arm their practice was assigned to, and all physicians were not informed of the frequency at which patient status reports would be delivered for either arm" (p. 867 Yeung 2012)
Incomplete outcome data (attrition bias) All outcomes	High risk	Per protocol analysis performed - 273/915 (29.9%) participants not included. Imbalance in numbers excluded between groups
Selective reporting (reporting bias)	High risk	The CGI-S and PGI-S were both recorded at baseline and different time points throughout the study but not reported. They were not used as interventions but should have been reported as outcomes. See p. 867 Yeung 2012 for full details No response was forthcoming from the authors on enquiring about these data
Other bias	Low risk	No other perceived bias

De Jong 2012

Methods	**Study design:** Individual randomised controlled trial **Setting:** 3 outpatient clinics in 2 mental health care institutions **Country:** Netherlands
Participants	**Diagnosis:** • 24% Mood disorder • 22% Adjustment disorder • 22% Anxiety disorder • 7% Personality disorder • 3% Eating disorder • 3% Diagnosed in childhood • 2% Substance related • 2% Somatoform disorder • 2% Impulse control disorder

	• 11% Other diagnoses **Method of diagnosis:** Routine diagnosis by the treating therapist, using the DSM-IV **Age:** Mean age 36.8, SD 12.0, range not given **Sex:** 61% female **Number:** 544 randomised, (feedback group 269, controls 275), of whom 413 (76%) were followed up (feedback group 206, controls 207) **Inclusion criteria** • Attending OP psychotherapy clinic for a minimum of 3 sessions and willing to complete outcome measures **Exclusion criteria** • Fewer than 3 outcome questionnaire administrations • Psychotic disorder • Mental retardation • Current crisis at time of referral • Non-verbal treatment • Group therapy as main treatment • Re-referral within same treatment centre within 6 months • Insufficient command of Dutch **Co-morbidities** • 8% Personality disorder • 37% Multiple Axis 1 disorders • 24% Comorbidity Axis 1 and 2 **Losses to follow-up/withdrawal:** 131 failed to complete baseline or 3 follow-up outcome measures Feedback group: < 3 sessions of treatment (24), < 33% OQ-45 administration (21) , stopped completing OQ-45 questionnaires before session 3 (13), baseline OQ-45 missing (5) Control group: < 3 sessions of treatment (30), < 33% OQ-45 administration (17), stopped completing OQ-45 questionnaires before session 3 (17), baseline OQ-45 missing (4) **Demographics considered:** Marital status, education **Ethnicity:** Not stated
Interventions	**PROM used as intervention**: Outcome Questionnaire 45 (OQ-45) Participants were randomly assigned to either **1) Feedback to therapist** **Duration:** variable, number of clinic visits determined by therapist Therapist given feedback prior to seeing client. (Therapist free to discuss feedback with participant) **2) Control group** **Duration:** variable, number of clinic visits determined by therapist The comparison group also completed OQ-45s but their scores were kept hidden from the therapists and participants
Outcomes	**Time points for assessment:** collected at every visit for first 5 visits, then every 5 visits for a year and at last clinic visit **Outcomes of the trial** (as reported): • Rate of change of OQ-45 total score* • Reliable change (change >14 points)

	Subgroups: Intervention and control groups were sub-divided into 'on-track (OT)', and 'not on track (NOT)', and rate-of-change results reported separately for two subgroups *outcomes prespecified for this review
Notes	**Dates of study:** Not stated **Funding:** Not stated

Risk of bias

Bias	**Authors' judgement**	**Support for judgement**
Random sequence generation (selection bias)	Low risk	Author reported that participants were assigned completely at random to feedback or no feedback using the feedback software
Allocation concealment (selection bias)	Low risk	Author reported that participants were not aware of their condition, unless therapists in the feedback group decided to discuss the feedback with the participant - this was explicitly allowed
Blinding of participants and personnel (performance bias) All outcomes	High risk	Not possible to blind clinicians and study personnel due to the nature of the intervention
Blinding of outcome assessment (detection bias) All outcomes	High risk	The PROM used for feedback was also used for outcome assessment, so the participants themselves were the outcome assessors and they were not blind to whether or not they received the intervention
Incomplete outcome data (attrition bias) All outcomes	High risk	High rates of attrition: • Feedback group 63/269 (23%) did not complete 3 PROMs • Control group 68/275 (25%) did not complete 3 PROMs
Selective reporting (reporting bias)	Low risk	No evidence of selective reporting found
Other bias	Low risk	No other perceived bias

De Jong 2014

Methods	**Study design:** Individual randomised controlled trial **Setting:** Outpatient clinics in mental health care institutions or private practices **Country:** Netherlands
Participants	**Diagnosis:** • 27% Mood disorder • 18% Adjustment disorder • 10% Anxiety disorder • 14% Relational problems (V codes) • 18% Other diagnoses • 13% No diagnosis **Method of diagnosis:** Routine diagnosis by the treating therapist, using the DSM-IV **Age:** Mean age 38.2, SD 12.0, range not given **Sex:** 68% female **Number:** 604 randomised, (therapist feedback only 205, therapist and patient feedback 207, controls 192), of whom 475 (79%) were followed up (therapist feedback only 159, therapist and patient feedback 172, controls 144) **Inclusion criteria:** • Attending OP psychotherapy clinic for a minimum of three sessions and willing to complete outcome measures **Exclusion criteria:** • Fewer than three outcome questionnaire administrations **Co-morbidities:** • 39% Personality disorder • 46% Comorbidity within axis 1 • 37% Comorbidity axes 1 and 2 **Losses to follow-up/withdrawal:** 129 failed to complete the three outcome measures **Demographics considered:** Education beyond high school **Ethnicity:** Not stated
Interventions	**PROM used as intervention**: Outcome Questionnaire 45 (OQ-45) and message comparing current OQ-45 total score, baseline score, and cut-off score for normal functioning Participants were randomly assigned to either: **1) Feedback to therapist only** **Duration:** variable, number of clinic visits determined by therapist Therapist given feedback prior to seeing client. (Therapist free to discuss feedback with participant) **2) Feedback to therapist and participant** **Duration:** variable, number of clinic visits determined by therapist. Participants received the same feedback as the therapists **3) Control group** **Duration:** variable, number of clinic visits determined by therapist The comparison group also completed OQ-45s but their scores were kept hidden from the therapists and participants
Outcomes	**Time points for assessment:** at last clinic visit. Analyses were conducted for subgroups of short-term (up to 35 weeks) and long-term (35-78 weeks) therapy **Outcomes of the trial** (as reported): Rate of change of OQ-45 total score*

	Subgroups: intervention and control groups were sub-divided into 'on-track (OT)', and 'not on track (NOT)', and rate-of-change results reported separately for NOT subgroups only *outcomes prespecified for this review
Notes	**Dates of study:** 1 July 2006- 30 June 2011 **Funding:** Netherlands Organization for Health Research and Development (ZonMW)

Risk of bias

Bias	**Authors' judgement**	**Support for judgement**
Random sequence generation (selection bias)	Low risk	"the online feedback system allocated the patient to one of the three conditions" The author reported that block randomisation was used to ensure each clinician had participants in all three conditions
Allocation concealment (selection bias)	Low risk	"the online feedback system allocated the patient to one of the three conditions"
Blinding of participants and personnel (performance bias) All outcomes	High risk	Not possible to blind participants and personnel due to the nature of the intervention
Blinding of outcome assessment (detection bias) All outcomes	High risk	The PROM used for feedback was also used for outcome assessment, so the participants themselves were the outcome assessors and they were not blind to whether or not they received the intervention
Incomplete outcome data (attrition bias) All outcomes	High risk	High rates of attrition: • Feedback to therapist group 46/205 (22%) did not complete 3 PROMs • Feedback to therapist and participant group 35/207 (17%) did not complete 3 PROMs • Control group 48/192 (25%) did not complete 3 PROMs
Selective reporting (reporting bias)	Low risk	No evidence of selective reporting found
Other bias	Low risk	No other perceived bias

Hansson 2013

Methods	**Study design:** Individual randomised controlled trial **Setting:** Psychiatric outpatients: 2 sites in Malmö **Country:** Sweden
Participants	**Diagnosis:** • Depression 119 (32%) • Bipolar disorder 29 (8%) • Anxiety syndrome 94 (25%) • Personality disorder 45 (12%) • Other diagnoses 14% • Missing/no diagnosis 32 (9%) **Method of diagnosis:** Not specified **Age:** Mean (SD): 38 (12.8) intervention group, 39 (14.1) control group **Sex:** 274 (73%) female, 100 (27%) male **Number:** 374 randomised (188 intervention group, 186 control group), all in ITT analysis; 262 followed up and in per protocol analysis (136 intervention group, 126 control group) **Inclusion criteria:** • Clinic attenders with mental disorders **Exclusion criteria:** • Substance use disorders • Schizophrenia • Other psychotic disorders **Co-morbidities:** Not stated **Losses to follow-up/withdrawals:** • Intervention group 52/188 (28%) did not complete follow up PROM • Control group 60/186 (32%) did not complete follow-up PROM **Demographics considered:** • Employment status • Marital status **Ethnicity:** • 32/238 (13%) with Social Insurance System data not born in Sweden
Interventions	**PROM used as intervention**: Outcome Questionnaire 45 (OQ-45) with feedback to both therapist and client Participants were randomly assigned to either: **1) Intervention group (188)** **Duration:** variable, number of clinic visits determined by therapist Therapist received a feedback message showing total score on OQ-45, the subscales and a diagram of treatment progress. Therapist could read feedback prior to seeing participant. Participant received feedback via treatment progress diagram **2) Control group (186)** **Duration:** variable, number of clinic visits determined by therapist Participants completed OQ-45 but no feedback to clinician or participant
Outcomes	**Time points for assessment:** Each clinic visit, reported at last clinic visit **Outcomes of the trial** (as reported): • Total OQ-45 scores* • Symptom distress sub scale scores* • Interpersonal difficulties sub scale scores*

	• Social function sub scale scores* • Frequency of OQ-45 scores representing alert status *outcomes prespecified for this review
Notes	**Dates of study:** 12 February 2007 to 10 February 2008 **Source of funding:** A grant from the Improved process for reporting of illness in Skåne, Skåne County Council; the Skåne County Council's Research and Development Foundation and the Swedish Social Insurance Agency, Malmö

Risk of bias

Bias	**Authors' judgement**	**Support for judgement**
Random sequence generation (selection bias)	Low risk	Quote: "Two different versions (feedback and control) of patient information were put into envelopes in a pre-randomized order. The randomization list was prepared using a computer program, which assigned the patient to one of the two groups at random"
Allocation concealment (selection bias)	Low risk	Quote: "The sealed envelopes were available at the reception and handed out in the same order as the patients were registered." "Everyone involved - patient, receptionist, therapist and researcher - were blinded to the allocation"
Blinding of participants and personnel (performance bias) All outcomes	High risk	Not possible to blind participants and personnel due to the nature of the intervention
Blinding of outcome assessment (detection bias) All outcomes	High risk	The PROM used for feedback was also used for outcome assessment, so the participants themselves were the outcome assessors and they were not blind to whether or not they received the intervention
Incomplete outcome data (attrition bias) All outcomes	High risk	High rates of attrition: • Intervention group 52/188 (28%) did not complete follow up PROM • Control group 60/186 (32%) did not complete follow-up PROM However, ITT analysis performed as well as per protocol
Selective reporting (reporting bias)	Low risk	ITT analysis performed as well as per protocol

Other bias	Low risk	No other perceived bias

Hawkins 2004

Methods	**Study design:** Individual randomised controlled trial **Setting:** Psychotherapy clinic **Country:** USA
Participants	**Diagnosis:** Axis I mood disorders (74%) and anxiety disorders (21%); 65 (32%) received two diagnoses **Method of diagnosis:** Routine diagnosis by the treating therapist **Age:** Mean age 30.8, SD 10.5, range not given. **Sex:** 137 female, 64 male, 112 not reported **Number:** 313 randomised, of whom 201 were followed up (therapist feedback only 70, therapist and client feedback 67, controls 64) **Inclusion criteria:** • Attending OP psychotherapy clinic for a minimum of 2 sessions and willing to complete outcome measures **Exclusion criteria:** • Failure to attend for a second session • Prescribed new medications or a change in medications during treatment **Co-morbidities:** Not stated **Losses to follow-up/withdrawal:** 108 failed to attend for a second session, 1 was removed by the therapist because it was thought the feedback was potentially detrimental, and 3 declined to complete the outcome measure and removed themselves from the study **Demographics considered:** Marital status, employment status **Ethnicity:** • 190 (94%) white • 3 (1.5%) African American • 3 (1.5%) Hispanic/Latino • 2 (1%) Asian American • 3 (1.5%) Pacific Islander/other
Interventions	**PROM used as intervention**: Outcome Questionnaire 45 (OQ-45) and algorithm on recommended actions as a function of number of treatment sessions provided and level of distress; 13 different instructions for therapists and 9 for clients Participants were randomly assigned to either: **1) Feedback to therapist only** **Duration:** variable, number of clinic visits determined by therapist Therapist received feedback that included 4 colour codes with actions recommended for each (white: consider termination; green: no change; yellow: consider altering treatment; red: review and decide on new course of action). Therapist given feedback prior to seeing client **2) Feedback to therapist and client** **Duration:** variable, number of clinic visits determined by therapist **3) Control group** **Duration:** variable, number of clinic visits determined by therapist

	The comparison group also completed OQ-45s but their scores were kept hidden from the therapists and clients
Outcomes	**Time points for assessment:** at last clinic visit **Outcomes of the trial** (as reported): • OQ-45 total score* • Proportion with reliable change in score (14+) reported for 'NOT' subgroup only • Proportion with clinically significant change in score (to below 64/180) reported for 'NOT' subgroup only • Outcome of potential treatment non responders • Effect of feedback on amount of psychotherapy **Subgroups**: Both intervention and control groups were sub-divided into 'on-track (OT)', i.e. green or white coded, and 'not on track (NOT)', i.e. yellow or red coded, and extent of change results reported separately for NOT subgroup only *outcomes prespecified for this review
Notes	**Dates of study:** Not stated **Source of funding:** Not stated

Risk of bias

Bias	Authors' judgement	Support for judgement
Random sequence generation (selection bias)	Low risk	Quote: "Patients were assigned to treatment conditions using a randomized block design, with therapists serving as the blocking variable"
Allocation concealment (selection bias)	Unclear risk	No details available, no response to enquiry to author
Blinding of participants and personnel (performance bias) All outcomes	High risk	Not possible to blind participants and personnel due to the nature of the intervention
Blinding of outcome assessment (detection bias) All outcomes	High risk	The PROM used for feedback was also used for outcome assessment, so the participants themselves were the outcome assessors and they were not blind to whether or not they received the intervention
Incomplete outcome data (attrition bias) All outcomes	High risk	112/313 participants (35.8%) were excluded from the analysis (108 did not attend at least one follow-up session after initial assessment, 3 did not complete the outcome measures, and 1 was removed by their therapist)

Hawkins 2004 (*Continued*)

Selective reporting (reporting bias)	Unclear risk	Reliable change and clinically significant change results not reported for 'on-track (OT)' subgroup
Other bias	Unclear risk	Clients started on medication, or receiving a change in medication, during treatment were excluded

Lambert 2001

Methods	**Study design:** Individual randomised controlled trial **Setting:** University counselling centre **Country:** USA
Participants	**Diagnosis:** 80% diagnosed, of whom: • 27% mood disorder • 14% adjustment disorder • 9% anxiety disorder • 5% somatoform disorder • 19% V-code diagnosis • 26% a variety of other disorders • 20% undiagnosed **Method of diagnosis:** Routine diagnosis by the treating clinician **Age:** Mean 22.23 years, range 17-57 **Sex:** 427 female, 183 male **Number:** 609 randomised (307 intervention group, 302 control group), all followed up and in per protocol analysis **Inclusion criteria:** • Consecutive centre clients who had at least one follow-up appointment **Exclusion criteria:** • None stated **Co-morbidities:** Not stated **Losses to follow-up/withdrawa**l: None reported **Demographics considered:** Not stated **Ethnicity:** • 88% white • 4% Hispanic • 3% Pacific Islander/Asian • 5% mixed
Interventions	**PROM used as intervention**: Outcome Questionnaire 45 (OQ-45) and algorithm on recommended actions. Feedback to clinician (but could be shared with client and was in some cases at least by all but 6 therapists) Participants were randomly assigned to either: **1) Intervention group (307)** **Duration:** variable, number of clinic visits determined by therapist Therapist received feedback that included 4 colour codes with actions recommended for each (white: consider termination; green: no change; yellow: consider altering treatment;

	red: review and decide on new course of action) Therapist given feedback prior to seeing client **2) Control group (302)** **Duration:** variable, number of clinic visits determined by therapist The comparison group also completed OQ-45s but their scores were kept hidden from the therapists and clients
Outcomes	**Time points for assessment:** collected at baseline, weekly and at last clinic visit **Outcomes of the trial** (as reported): • OQ-45 total score* • Proportion with reliable change in score (14+) • Proportion with clinically significant change in score (to below 64/180) • Effect of feedback on amount of psychotherapy • Exploratory analyses of timing of feedback • Assessment of therapist experience of recipient of feedback **Subgroups:** Both intervention and control groups were sub-divided into 'on-track (OT)', i.e. green or white coded, and 'not on track (NOT)', i.e. yellow or red coded, and results reported separately for each subgroup within intervention and control arms *outcomes prespecified for this review
Notes	**Dates of study:** Enrolment from October 1998-April 1999 **Source of funding:** University funded

Risk of bias

Bias	**Authors' judgement**	**Support for judgement**
Random sequence generation (selection bias)	Unclear risk	Quote: "Approximately half (n = 307) were randomly assigned to the experimental (feedback) group and half (n = 302) were randomly assigned to the control (no feedback) group'
Allocation concealment (selection bias)	Unclear risk	Quote: 'Approximately half (n = 307) were randomly assigned to the experimental (feedback) group and half (n = 302) were randomly assigned to the control (no feedback) group'
Blinding of participants and personnel (performance bias) All outcomes	High risk	Not possible to blind participants and personnel due to the nature of the intervention
Blinding of outcome assessment (detection bias) All outcomes	High risk	The PROM used for feedback was also used for outcome assessment, so the participants themselves were the outcome assessors and they were not blind to whether or not they received the intervention

Lambert 2001 (*Continued*)

Incomplete outcome data (attrition bias) All outcomes	Low risk	According to the paper, 609 were randomised and all were included in the analysis without a single dropout (we were unable to confirm this with the author)
Selective reporting (reporting bias)	High risk	Clinically significant change details not provided for on-track participants (majority of participants)
Other bias	Low risk	No other perceived bias

Mathias 1994

Methods	**Study design:** Cluster randomised controlled trial **Setting:** Primary care (Health Maintenance Organisation (HMO)) **Country:** USA
Participants	**Diagnosis:** **Method of diagnosis:** Diagnostic Interview Schedule (DIS) for the DSM-III-R **Age:** Mean 42 yrs (SD 10) in intervention group; 44 (11) in controls. Range 21-65 **Sex:** 336 female, 237 male, 45 not reported **Number:** 618 randomised (389 intervention, 229 control) **Inclusion criteria:** Symptoms of anxiety and depression on Hopkins Symptom Checklist (SCL-90) above 'threshold' on two occasions **Exclusion criteria:** Previously diagnosed mental health condition or received treatment in the past 6 months **Co-morbidities:** 394 (69%) had co-morbidities; not specified further **Losses to follow-up/withdrawals:** • 45 (7.3%) dropped out: (32 (8.2%) intervention, 13 (5.7%) control) **Demographics considered:** • Gender • Age • Education • Income • Marital status. **Ethnicity:** 112 non-white
Interventions	**PROM used as intervention**: Mental Health Patient Profile, constructed from SCL-90, DIS, and SF-36 Participants were randomly assigned to either: **1) Intervention group (389):** The PROMs were administered by researchers outside the practice and the results summarised for the treating physicians. Feedback to clinician only **2) Control group (229):** No feedback of PROM scores to clinician or participant
Outcomes	**Outcomes:** **Mathias 1994:** • Global anxiety score (GAS)*

	• Global severity index (GSI) • Highest Anxiety Subscale Score (HASS) (GAS, GSI, and HASS all derived from SCL-90) • SF-36 (nine subscale scores)* **Mazonson 1996:** • Chart notation of anxiety, depression or other mental health diagnosis or symptoms • Referral to mental health specialist* • Prescription of psychotropic medications* • Hospitalisation • Clinic visits **Subgroups**: 4 severity subgroups (anxiety symptoms only, anxiety symptoms and disorder, anxiety and depression symptoms, anxiety and depression disorders) **Time points for assessment:** 12 weeks and 5 months *outcomes prespecified for this review
Notes	**Duration:** Not stated **Funding:** Supported by a grant from the Upjohn Company, and Take Care Colorado Characteristics and data obtained from both Mathias 1994 (primary reference) and Mazonson 1996

Risk of bias

Bias	Authors' judgement	Support for judgement
Random sequence generation (selection bias)	High risk	Quote: "The physicians were randomized by call group to either the demonstration or control arm." Baseline imbalances in participant numbers and demographics
Allocation concealment (selection bias)	High risk	Cluster randomisation means physicians were aware of allocation
Blinding of participants and personnel (performance bias) All outcomes	High risk	Not possible to blind participants and personnel due to the nature of the intervention
Blinding of outcome assessment (detection bias) All outcomes	High risk	Outcome assessment was carried out by the researchers administering the PROMs and feeding back the results to the physicians, who were therefore aware of allocation
Incomplete outcome data (attrition bias) All outcomes	High risk	The 45 participants lost to follow-up had higher mean scores for SF-36 than the participants followed up
Selective reporting (reporting bias)	Low risk	No evidence of selective reporting found

Mathias 1994 (*Continued*)

Other bias	Low risk	No other perceived bias

Murphy 2012

Methods	**Study design:** Individual randomised controlled trial **Setting:** University counselling service **Country:** Ireland
Participants	**Diagnosis:** 4 'dominant representations': • Anxiety (29.1%) • Depression (29.1%) • Relationships (19.1%) • Other (22.7%). **Method of diagnosis:** Routine diagnosis by the treating clinician **Age:** Mean (SD) = 23.82 (6.46) years, range 18-59 **Sex:** 58.2% female: 50.8% in intervention group and 66.7% in control group **Number:** 149 randomised at clinic intake, of which 110 followed up and included in per protocol analysis (59 intervention group and 51 control) **Inclusion criteria:** • Consecutive centre clients who had at least 1 follow-up appointment **Exclusion criteria:** • Attending for an emergency drop-in appointment (where the client was in high distress) • Attending a scheduled screening for accessing online support **Co-morbidities: n**ot stated **Losses to follow-up/withdrawa**l 32/180 originally assigned declined consent or dropped out before the first assessment. A further 39/149 (26.2%) failed to complete the second assessment (18 intervention group, 21 control group) **Demographics considered:** Not stated **Ethnicity:** Not stated
Interventions	**PROM used as intervention**: Outcome rating scale (ORS) scores. Feedback to both clinician and participant who reviewed the scores together Participants were randomly assigned to either: **1) Intervention group (59)** **Duration:** variable, number of clinic visits determined by therapist Therapist received feedback that included a graph of projected progress, and actual progress. Therapist given feedback while seeing client **2) Control group (51)** **Duration:** variable, number of clinic visits determined by therapist The comparison group also completed ORS but their scores were kept hidden from the therapists and clients
Outcomes	**Time points for assessment:** at last clinic visit **Outcomes of the trial** (as reported): • ORS total score* • Proportion with reliable change in score (>5), collectively and per diagnosis • Length of treatment

	Subgroups: Both intervention and control groups were sub-divided into 4 diagnostic groups (above) *outcomes prespecified for this review
Notes	**Dates of study:** Enrolment from November 2008-February 2009 **Source of funding:** University funded

Risk of bias

Bias	**Authors' judgement**	**Support for judgement**
Random sequence generation (selection bias)	Low risk	Quote: "For randomisation an online random number generator was utilised"
Allocation concealment (selection bias)	Low risk	Quote: "Clients were randomly assigned at intake to either the 'feedback' or 'no feedback' condition" (before assignment to a therapist)
Blinding of participants and personnel (performance bias) All outcomes	High risk	Not possible to blind participants and personnel due to the nature of the intervention
Blinding of outcome assessment (detection bias) All outcomes	High risk	The PROM used for feedback was also used for outcome assessment, so the participants themselves were the outcome assessors and they were not blind to whether or not they received the intervention
Incomplete outcome data (attrition bias) All outcomes	High risk	180 originally assigned in total; 31 did not complete first assessment; 149 randomised; 39 (26.2%) failed to complete the second assessment; 110/149 included in per protocol analysis
Selective reporting (reporting bias)	Unclear risk	Proportion with clinically significant change in ORS score (to 25+/40) not reported
Other bias	Low risk	No other perceived bias

Probst 2013

Methods	**Study design:** Individual randomised controlled trial **Setting:** Inpatient facility for people with psychosomatic disorder **Country:** Germany
Participants	**Diagnosis:** Psychosomatic disorder **Method of diagnosis:** Not stated **Age:** Mean (SD) = 47.62 (13.44) years, range not stated (participants in per protocol analysis) **Sex:** 60.6% female: 264/436 participants **Number:** 436 randomised, of which 252 followed up and included in per protocol analysis (20 intervention group and 23 control of the 'not on track' group, 111 intervention group and 98 control group of the 'on track' group) **Inclusion criteria:** • Inpatient with a psychosomatic disorder **Exclusion criteria:** • No baseline OQ-45 completed **Co-morbidities:** 'Not on track' group: • Depressive disorders (76.7%): 78.3% of intervention group, 75.0% of control group • Somatoform disorders (58.1%): 52.2% of intervention group, 65.0% of control group • Anxiety disorders (20.9%): 30.4% of intervention group, 10.0% of control group 'On track' group: • Depressive disorders (64.6%): 62.2% of intervention group, 67.4% of control group • Somatoform disorders (58.9%): 58.6% of intervention group, 59.2% of control group • Anxiety disorders (26.3%): 23.4% of intervention group, 29.6% of control group **Losses to follow-up/withdrawal:** 184/436 originally randomised excluded as treatment duration not long enough to provide OQ-45 data for the intake week and at least 2 more weeks (69/184 due to severe distress, 31/184 data available for only 1 or 2 weeks, 84/184 unclear) 'On-track' group missing data: • Intervention group 37/111, no dropouts • Control group 29/98, no dropouts **Demographics considered:** • Education **Ethnicity:** Not stated
Interventions	**PROM used as intervention**: OQ-45 (German version), ASC (Assessment of Signal Cases). Feedback given to clinician only, but free to discuss with clients Participants were randomly assigned to either: **1) Intervention group (111 in 'on track' group, 23 in 'not on track' group)** **Duration:** Mean duration of treatment 3.6 weeks OQ-45 scores and ASC data of each participant given to therapists on a weekly basis, after being entered into OQ-Analyst. Therapist received feedback indicating if participant at risk of deterioration: yellow: consider altering treatment; red: review and decide on new course of action). Therapist given feedback prior to seeing client **2) Control group (98 in 'on track' group, 20 in 'not on track' group)**

	Duration: Mean duration of treatment 3.4 weeks OQ-45 and ASC completed by each participant every week but not shared with therapists
Outcomes	**Time points for assessment:** (5), Baseline, weeks 1, 2, 3 and discharge week or last available OQ-45 assessment **Outcomes of the trial** (as reported): • Mean change in OQ-45 total score* • Mean change in OQ-45 symptom distress scale • Mean change in OQ-45 interpersonal relations scale • Mean change in OQ-45 social performance scale* • Mean change from baseline OQ-45 score to last measurement point in 'not on track' group* • Reliable change index (RCI) *outcomes prespecified for this review
Notes	**Dates of study:** October 2010-July 2012 **Sources of funding:** University Professorship awarded to Michael Lambert Characteristics and data obtained from both Probst 2013 (primary reference) and Probst 2014

Risk of bias

Bias	Authors' judgement	Support for judgement
Random sequence generation (selection bias)	Unclear risk	No details provided. No response from authors when contacted to clarify
Allocation concealment (selection bias)	Unclear risk	No details provided
Blinding of participants and personnel (performance bias) All outcomes	High risk	Not possible to blind therapists due to the nature of the intervention
Blinding of outcome assessment (detection bias) All outcomes	High risk	The PROM used for feedback was also used for outcome assessment, so the participants themselves were the outcome assessors and they were not blind to whether or not they received the intervention
Incomplete outcome data (attrition bias) All outcomes	High risk	184/436 (42.2%) participants excluded as did not complete 1 or more assessments. Analysis done per protocol
Selective reporting (reporting bias)	Low risk	No evidence of selective reporting found
Other bias	Low risk	No other perceived bias

Reese 2009a

Methods	**Study design:** Individual randomised controlled trial (Study 1 of 2 described in paper) **Setting:** University counselling service **Country:** USA
Participants	**Diagnosis:** Not stated in paper. The lead author confirmed by email that more than 90% of the study participants would have had qualifying clinical diagnoses of anxiety and depressive disorders **Method of diagnosis:** Clinical only **Age:** Mean (SD) = 20.17 (1.9) years, range 18 -27 **Sex:** 53 female, 18 male, 60 not reported **Number:** 131 randomised at clinic intake, of which 74 followed up and included in per protocol analysis (50 intervention group and 24 control) **Inclusion criteria:** • All clients referred to the services who attended for at least one follow-up appointment **Exclusion criteria:** • Receiving couples therapy or family therapy **Co-morbidities:** Not stated **Losses to follow-up/withdrawal:** 57/131 (43.5%) either failed to return for a second session (24), did not complete the PROMs consistently in the feedback arm (5) or did not complete a post-treatment measure in the no-feedback arm (33) **Demographics considered:** Not stated **Ethnicity:** • 78.4% white • 4.1% African American • 2.7% Asian American • 6.8% Hispanic/Latino • 5.4% 'international students'
Interventions	**PROM used as intervention**: PCOMS (Partners for Change Outcome Management System) including ORS (Outcome Rating Scale) and SRS (Session Rating Scale) Feedback to both clinician and participant who reviewed the scores together Participants were randomly assigned to either: **1) Intervention group** (50) **Duration:** variable, number of clinic visits determined by therapist Therapists received and viewed feedback together with clients **2) Control group** (24) **Duration:** variable, number of clinic visits determined by therapist The comparison group also completed ORS but their scores were kept hidden from the therapists and clients
Outcomes	**Time points for assessment:** baseline, weekly, at last clinic visit **Outcomes of the trial** (as reported): • ORS total score* • Proportion with reliable change in score (> 5) • Survival plots for achieving reliable change • Difference in number of treatment sessions between groups *outcomes prespecified for this review

Notes	**Dates of study:** Not stated **Source of funding:** Not stated	
Risk of bias		
Bias	**Authors' judgement**	**Support for judgement**
Random sequence generation (selection bias)	Low risk	Quote: ..."roughly half of the participants were originally randomly assigned to the feedback...via a randomised block design to help control for therapist effects". and "..... the first client was randomized using a random number generator to either the feedback or TAU condition.The second client was then assigned to the other condition." (Author correspondence, see Table 1)
Allocation concealment (selection bias)	Low risk	"This was done by the person who assigned clients at the respective centers, after enrollment into the study. Investigators and client participants could not foresee which condition a participant would placed into. " (Author correspondence)
Blinding of participants and personnel (performance bias) All outcomes	High risk	Not possible to blind participants and personnel due to the nature of the intervention
Blinding of outcome assessment (detection bias) All outcomes	High risk	The PROM used for feedback was also used for outcome assessment, so the participants themselves were the outcome assessors and they were not blind to whether or not they received the intervention
Incomplete outcome data (attrition bias) All outcomes	Unclear risk	Substantial proportion failed to complete and were left out of per protocol analysis (57/131, 43.5%)
Selective reporting (reporting bias)	High risk	SRS completed at end of each session in feedback group, but results not reported
Other bias	Unclear risk	No other perceived bias

Reese 2009b

Methods	**Study design:** Cluster randomised controlled trial (Study 2 of 2 studies described in paper) **Setting:** Community-based graduate (Masters) training clinic **Country:** USA
Participants	**Diagnosis:** Not stated in paper. The lead author confirmed by email that more than 90% of the study participants would have had qualifying clinical diagnoses of anxiety and depressive disorders **Method of diagnosis:** Clinical only **Age:** Mean (SD) = 32.96 (12.32) years, range 18-69 **Sex:** 51 female, 21 male, 24 not reported **Number:** 96 randomised at clinic intake, of which 74 followed up and included in per protocol analysis (45 intervention group and 29 control) **Inclusion criteria:** • All clients referred to the services who attended for at least one follow-up appointment **Exclusion criteria:** • Receiving couples or family therapy **Co-morbidities:** Not stated **Losses to follow-up/withdrawal:** 22/96 (22.9%) either failed to return for a second session (8), did not complete the PROMs consistently in the feedback arm (4) or did not complete a post-treatment measure in the no-feedback arm (10) **Demographics considered:** Not stated **Ethnicity:** • 79.6% white • 3.7% African American • 14.6% Hispanic/Latino • 2.1% undeclared
Interventions	**PROM used as intervention**: PCOMS (Partners for Change Outcome Management System) including ORS (Outcome Rating Scale) and SRS (Session Rating Scale) Feedback to both clinician and participant who reviewed the scores together Participants were randomly assigned to either: **1) Intervention group** (45) **Duration:** variable, number of clinic visits determined by therapist Therapists received and viewed feedback together with clients **2) Control group** (29) **Duration:** variable, number of clinic visits determined by therapist The comparison group also completed ORS but their scores were kept hidden from the therapists and clients
Outcomes	**Time points for assessment:** baseline, weekly, at last clinic visit **Outcomes of the trial** (as reported): • ORS total score* • Proportion with reliable change in score (>5) • Survival plots for achieving reliable change • Difference in number of treatment sessions between groups *outcomes prespecified for this review

Reese 2009b (*Continued*)

Notes	**Dates of study:** Not stated **Source of funding:** Not stated

Risk of bias

Bias	Authors' judgement	Support for judgement
Random sequence generation (selection bias)	Low risk	Quote:"Therapists, rather than clients were randomly assigned to the feedback and no-feedback conditions"
Allocation concealment (selection bias)	High risk	Cluster randomisation means therapists were aware of allocation
Blinding of participants and personnel (performance bias) All outcomes	High risk	Not possible to blind participants and personnel due to the nature of the intervention
Blinding of outcome assessment (detection bias) All outcomes	High risk	The PROM used for feedback was also used for outcome assessment, so the participants themselves were the outcome assessors and they were not blind to whether or not they received the intervention
Incomplete outcome data (attrition bias) All outcomes	Unclear risk	Substantial proportion failed to complete and were left out of per protocol analysis (22/96, 23%)
Selective reporting (reporting bias)	High risk	SRS completed at end of each session in feedback group, but results not reported
Other bias	Unclear risk	No other perceived bias

Scheidt 2012

Methods	**Study design:** Cluster randomised controlled trial **Setting:** Private outpatient psychotherapy **Country:** Germany
Participants	**Diagnosis:** Mean (SD) number of F-diagnoses according to ICD-10: Intervention group 1.71 (0.92), Control group 1.46 (0.77) **Method of diagnosis:** According to ICD (intervention group: ICDL-Checklist, control group: ICD -10 criteria) **Age:** Intervention group: mean 40.16 years, SD 11.38. Control group: mean 41.27 years, SD 11.03. Range not reported **Sex:** 1117 (68.6%) female (Wittmann et al., 2012) **Number:** 4452 approached, 1708 patients gave consent to participate: 1031 intervention group, 677 control group

	1629 randomised: 968 intervention group, 661 control (Wittmann et al. 2011) **Inclusion criteria:** Starting outpatient psychotherapy, between 1 April 2005 and 30 June 2010; diagnosis F3-F6, 18 years or older **Exclusion criteria:** Diagnosis F1 or F2 **Co-morbidities:** Not reported **Losses to follow-up/withdrawals:** 1598 (98.2%) completed baseline assessment, 597 (36.7%) completed post-treatment assessment, and 468 (28.8%) completed follow-up assessment 12 months post-treatment **Demographics considered:** Marital status, partnership status, persons living in household, living situation, education, training qualification, job status, income, ability to work **Ethnicity:** German nationality: intervention group 97.8%, control group 96.0%
Interventions	**PROM used as intervention**: Comprehensive inventory of psychometric measurement instruments. Decision rules ("reorientation of the expert system") were developed and optimised to guide decisions (on indications for and prolongation of psychotherapy) based on the feedback. No extra contact or treatment was given to the participants in the intervention group as a result of reorientation of the expert system Participants were randomly assigned to either: **1) Intervention group:** Feedback to clinician only on 4-point scale: consistent reduction of problems; reduction of problems; no clinically relevant changes; increase of problems **2) Control group:** No feedback to clinician or participant
Outcomes	**Outcomes:** • Brief Symptom Inventory BSI • Inventar für Interpersonale Probleme IIP-D • Secondary outcomes: Beck Depressionsinventar BDI; Fragebogen zu Körperbezogenen Ängsten, Kognitionen und Vermeidung AKV; Hamburger Zwangsinventar HZI; Eating Disorder Inventory EDI; Screening für Somatoforme Störungen SOMS (Questionnaires on body-related anxiety and cognitions); and Helping Alliance Questionnaire HAQ • Fragebogen zum Gesundheitszustand SF-12 **Time points for assessment:** Pre-treatment, post-treatment, and 12 months post-treatment
Notes	**Duration:** 01/05/2005-31/05/2011, 73 months **Funding:** Techniker Krankenkasse health insurance programme

Risk of bias

Bias	Authors' judgement	Support for judgement
Random sequence generation (selection bias)	Unclear risk	Drawing lots: "Aus jeder der Zellen wurden per Zufall die teilnehmenden Therapeuten zur IG oder KG zugelost" (Wittmann et al. , 2011, p42)

Scheidt 2012 (*Continued*)

Allocation concealment (selection bias)	High risk	Cluster randomisation means therapists were aware of allocation. Allocation was restricted according to gender and treatment modalities
Blinding of participants and personnel (performance bias) All outcomes	Unclear risk	Blinding of participants not reported.
Blinding of outcome assessment (detection bias) All outcomes	Unclear risk	Blinding of outcome assessors not reported
Incomplete outcome data (attrition bias) All outcomes	High risk	Only 36.7% completed post-treatment assessment, and 28.8% completed follow-up assessment 12 months post-treatment
Selective reporting (reporting bias)	Unclear risk	Unclear whether selective reporting had taken place
Other bias	Unclear risk	No other perceived bias

Simon 2012

Methods	**Study design:** Individual randomised controlled trial **Setting:** Hospital-based outpatient clinic **Country:** USA
Participants	**Diagnosis:** • 64% mood • 30% anxiety disorders • 5% substance abuse • 45.7% met criteria for two or more diagnoses **Method of diagnosis:** Routine diagnosis by the treating therapist **Age:** Mean age (SD) 36.10 (13.32), range not given **Sex:** 64.2% female (241), 34.9% male (129) of the 370 included in analysis, 94 not reported **Number:** 464 recruited, of whom 370 were followed up. 163 'on-track (OT)' all improved. 207 'not on track (NOT)' clients underwent 'Assessment for Signal Clients (ASC)' and were randomised to therapist and client feedback (109), or treatment as usual (98) **Inclusion criteria:** • Attending OP psychotherapy clinic for a minimum of 2 sessions and willing to complete outcome measures **Exclusion criteria:** • Failure to attend for a second session • Age < 18 • Exclusively receiving medication or forms of treatment other than individual

	psychotherapy **Co-morbidities:** Not stated **Losses to follow-up/withdrawal:** 94 failed to attend for a second session **Demographics considered:** • Marital status • Employment status **Ethnicity:** • 92.7% white • 1.9% African American • 2.4% Hispanic/Latino • 1.9% Asian American • 1.6% Pacific Islander or other
Interventions	**PROM used as intervention**: Outcome Questionnaire 45 (OQ-45), including 3 subscales: subjective discomfort, interpersonal relationships, and social role performance Participants were randomly assigned to either: **1) Feedback** **Duration:** variable, number of clinic visits determined by therapist Feedback to clinicians consisted of session-by-session OQ-45 progress feedback along with alerts to therapists each time a client took the measure. Therapists were given feedback prior to seeing clients that included four colour codes with actions recommended for each (white: consider termination; green: no change; yellow: consider altering treatment; red: review and decide on new course of action). Therapists were instructed to share OQ-45 scores with clients **Subgroups**: both intervention and control groups were sub-divided into 'on-track (OT)', i.e. green or white coded, and 'not on track (NOT)', i.e. yellow or red coded. The ASC was used for 'not on track (NOT)' clients only, and results were reported for the NOT subgroup only Therapists were also provided with a Clinical Support Tool (CST) intervention manual, which provided guidelines for interpreting the ASC, a decision tree, and an interventions list to prompt therapist action **2) Control group** **Duration:** variable, number of clinic visits determined by therapist The treatment-as-usual group also completed OQ-45s but their scores were kept hidden from the therapists and clients
Outcomes	**Time points for assessment:** Baseline, each clinic visit, at last clinic visit **Outcomes of the trial** (as reported): **All reported for NOT** subgroup**only:** • OQ-45 total scores* • Proportion with reliable change in score (14+) • Proportion with clinically significant change in score (to below 64/180) • Comparisons of scores per therapist • Effect of feedback on amount of psychotherapy *outcomes prespecified for this review
Notes	**Dates of study:** Not stated **Source of funding:** Susa Young Gates University Professorship awarded to Michael J Lambert

Simon 2012 (*Continued*)

Risk of bias		
Bias	**Authors' judgement**	**Support for judgement**
Random sequence generation (selection bias)	Low risk	Quote: "patients were randomly assigned by the research staff". "Patients in this study were randomly assigned to experimental conditions using a block randomized design, with therapists serving as the blocking variable." p. 640
Allocation concealment (selection bias)	Unclear risk	No details
Blinding of participants and personnel (performance bias) All outcomes	High risk	Not possible to blind participants and personnel due to the nature of the intervention
Blinding of outcome assessment (detection bias) All outcomes	High risk	The PROM used for feedback was also used for outcome assessment, so the participants themselves were the outcome assessors and they were not blind to whether or not they received the intervention
Incomplete outcome data (attrition bias) All outcomes	High risk	94/464 (20.3%) failed to complete the second assessment and were omitted from the analysis
Selective reporting (reporting bias)	Unclear risk	No results reported for OT subgroup, except that they all improved with no differences between feedback and no-feedback groups
Other bias	Low risk	No other perceived bias

Trudeau 2001

Methods	**Study design:** Individual randomised controlled trial **Setting:** Rural community mental health centre - 3 sites **Country:** USA
Participants	**Diagnosis:** • 67% Axis I disorder • 24% severe/recurrent Axis I disorder • < 2% psychotic disorders • < 1% substance abuse • 6% other **Method of diagnosis:** Not stated. Author confirmed by email that diagnosis was clinical, and that more than 90% would have had diagnoses of anxiety and depressive disorders

	Age: Control group mean age (SD) 37.5 (14.32), feedback group mean age (SD) 32.14 (10.51), non-feedback group mean age (SD) 32.91 (13.36) **Sex:** 72% female (91/127) **Number:** 127 (38 control, 66 feedback, 23 no feedback) **Inclusion criteria:** • People presenting for mental health therapy **Exclusion criteria:** • Not consenting to study **Co-morbidities:** Not stated **Losses to follow-up/withdrawal:** 14/38 in control group, 26/66 in feedback group and 10/23 in non-feedback group withdrew after T1. A further 18 withdrew after T2 (group status not given). No reasons given for drop-outs **Demographics considered:** • Education • Employment status • Income • Marital status • Family size **Ethnicity:** • 97% white • Remainder unreported
Interventions	**PROM used as intervention**: Outcome Questionnaire (OQ-45) with subscales 3 groups: 1) a **feedback** condition, in which the clients completed the OQ at each session, and the clinicians were provided with information regarding client progress following each session 2) a **non-feedback** condition in which clients completed the OQ at each session, but the clinicians were not provided with the results of the measures 3) a **control** condition in which clients were not assessed with the OQ measures at each session
Outcomes	**Time points for assessment:** (3) Baseline, 2 months, 4 months **Outcomes as reported by study authors:** • OQ scores* † • Total Mental Health score* • RAND health survey • Work/school questionnaire • AABH Service Utiilization questionnaire • AABH patient satisfaction questionnaire • Clinician evaluation of managed care *outcomes prespecified for this review † = no usable data provided for inclusion in this review
Notes	**Dates of study:** Not stated **Source of funding:** Study conducted for doctoral thesis, no source of funding disclosed
Risk of bias	

Trudeau 2001 (*Continued*)

Bias	Authors' judgement	Support for judgement
Random sequence generation (selection bias)	Low risk	"Clients were randomly assigned by case number to either the control condition for case numbers ending in 3, 6 or 9"
Allocation concealment (selection bias)	Unclear risk	No details provided
Blinding of participants and personnel (performance bias) All outcomes	High risk	Not possible to blind participants and personnel due to the nature of the intervention
Blinding of outcome assessment (detection bias) All outcomes	High risk	The PROM used for feedback was also used for outcome assessment, so the participants themselves were the outcome assessors and they were not blind to whether or not they received the intervention
Incomplete outcome data (attrition bias) All outcomes	High risk	14/38 dropped out in control group, 26/66 dropped out in feedback group and 10/23 dropped out in no feedback group. Further 18 across groups dropped out between T2 and T3. Balanced drop out rates but high, and substantially higher participants in feedback group compared to no feedback group
Selective reporting (reporting bias)	Low risk	No evidence of selective reporting
Other bias	Low risk	No other perceived bias

Whipple 2003

Methods	**Study design:** Individual randomised controlled trial **Setting:** University counselling centre **Country:** USA
Participants	**Diagnosis:** 74.6% diagnosed: • 35% 'V code' diagnosis • 29.2% mood disorder • 12.4% adjustment disorder • 10.1% anxiety disorder • 7% eating disorder • 6.3% 'other' **Method of diagnosis:** Routine diagnosis by the treating clinician **Age:** Mean 22.88 (SD 3.54), range 18-54 **Sex:** 648 female, 333 male, 358 not reported

	Number: 1339 randomised, of whom 981 (73.2%) followed up and in per protocol analysis (499 intervention, 482 control) **Inclusion criteria:** • Consecutive centre clients who had at least one follow-up appointment **Exclusion criteria:** • None stated **Co-morbidities:** Not stated **Losses to follow-up/withdrawa**l 1339 included originally, of whom 358 (26.7%) excluded due to not completing an outcome measure, or not returning for a second session **Demographics considered:** Not stated **Ethnicity:** • 86% white • 4.8% Hispanic • 2.1% Pacific Islander/Asian • 0.6% African American • 6.5% other or mixed
Interventions	**PROM used as intervention**: Outcome Questionnaire 45 (OQ-45) and algorithm on recommended actions. Feedback to clinician (but could be shared with client) Participants were randomly assigned to either: **1) Intervention group (499)** **Duration:** variable, number of clinic visits determined by therapist Therapist received feedback that included 4 colour codes with actions recommended for each (white: consider termination; green: no change; yellow: consider altering treatment; red: review and decide on new course of action). Therapist given feedback prior to seeing client. When clients identified as 'not on track (NOT)', therapists had option of using a clinical support tool (CST) which included a decision tree and a list of possible interventions **2) Control group (482)** **Duration:** variable, number of clinic visits determined by therapist The comparison group also completed OQ-45s but their scores were kept hidden from the therapists and clients
Outcomes	**Time points for assessment:** Baseline, per session and at last clinic visit **Outcomes of the trial** (as reported): • OQ-45 total score* • Proportion with clinically significant change in score (to below 64/180) • Differences in treatment length • Therapist effect on outcome **Subgroups:** both intervention and control groups were sub-divided into 'on-track (OT)', i.e. green or white coded, and 'not on track (NOT)', i.e. yellow or red coded, and results reported separately for each subgroup within intervention and control arms *outcomes prespecified for this review
Notes	**Dates of study:** Not stated **Source of funding:** Not stated
Risk of bias	

Whipple 2003 (*Continued*)

Bias	Authors' judgement	Support for judgement
Random sequence generation (selection bias)	Unclear risk	Quote: "The participants in the experimental (Fb) and control groups (NFb) were divided into groups based on random assignment"
Allocation concealment (selection bias)	Unclear risk	Not stated
Blinding of participants and personnel (performance bias) All outcomes	High risk	Not possible to blind participants and personnel due to the nature of the intervention
Blinding of outcome assessment (detection bias) All outcomes	High risk	The PROM used for feedback was also used for outcome assessment, so the participants themselves were the outcome assessors and they were not blind to whether or not they received the intervention
Incomplete outcome data (attrition bias) All outcomes	High risk	1339 randomised, of whom 358 (26.7%) excluded due to not completing an outcome measure, or not returning for a second session. Per protocol analysis undertaken
Selective reporting (reporting bias)	Low risk	No evidence of selective reporting found
Other bias	Low risk	No other perceived bias

Characteristics of excluded studies *[ordered by study ID]*

Study	Reason for exclusion
Anker 2009	Ineligible population
Brodey 2005	PROMs not used for outcome monitoring
Brody 1990	PROMs not used for outcome monitoring
Dobscha 2006	Complex quality improvement programme
Fluckiger 2012	PROMs not used for outcome monitoring
Haderlie 2012	Non randomised study

(*Continued*)

Harmon 2007	Non randomised study
Lambert 2002	Non randomised study
Newnham 2010	Non randomised study
Pedersen 2014	PROMs not used for outcome monitoring
Priebe 2007	Ineligible population
Puschner 2009	Ineligible population
Reese 2010	Ineligible population
Reese 2013	PROMs not used for outcome monitoring
Reeves 2010	Non randomised study
Rise 2012	PROMs not used for outcome monitoring (only therapeutic alliance and patient satisfaction reported)
Simon 2000	Complex quality improvement programme
Slade 2006	Ineligible population
Slade 2008	Non randomised study

Characteristics of studies awaiting assessment *[ordered by study ID]*

Gibbons 2015

Methods	Randomised controlled trial with 2 arms. Patients individually randomised to 8 weeks of individual therapy with either a clinician receiving weekly feedback reports, or a clinician not receiving weekly feedback reports
Participants	People seeking services for depression at a community mental health centre (CMHC) in Philadelphia, USA. Inclusion criterion: a clinically meaningful level of depressive symptoms (score of 11 or above on Inventory for Depressive Symptomatology, QIDS). Pre-dominantly female African-Americans, with a mean age of 39
Interventions	Community Clinician Feedback System (CCFS) including a clinical feedback report identifying patients who were not progressing as expected judged on basis of scores on BASIS-24 (24-Item Behavior and Symptom Identification Scale). Scores presented on coloured graph showing line of expected recovery along with patient's actual BASIS-24 scores. Patients 'off track' for improvement completed Community Clinician Feedback Questionnaire (CCFQ) covering demographic background, treatment motivation, attitudes and expectations about treatment, therapeutic alliance, suicide risk, substance use, perceived social support, psychosocial stressors, violence potential, personality disorder, interpersonal distress, interpersonal patterns, cognitive distortions, compensatory skills, and trauma history

Gibbons 2015 (*Continued*)

Outcomes	Primary outcome: rate of change across treatment weeks on the total score of the BASIS-24, plus percentage of patients achieving reliable change, clinically significant change, and both reliable and clinically significant change. Secondary outcomes: patient and clinician satisfaction
Notes	

Rise 2016

Methods	Open, individually randomised parallel-group controlled trial
Participants	Outpatients attending a mental health hospital in Norway. All patients offered treatment at the out-patient unit between 6 weeks and 3 months after referral were invited
Interventions	Partners for Change Outcome Management System (PCOMS) feedback scales, including the Outcome Rating Scale (ORS), and Session rating Scale (SRS)
Outcomes	Primary outcome measures: Behaviour and Symptom Identification Scale 32 (BASIS-32) and Patient Activation Measure (PAM) Secondary outcome measures: Treatment Alliance Scale (TAS), Client Satisfaction Questionnaire-8 (CSQ), Short Form-12 (SF-12), Outcome Rating Scale (ORS) and Session Rating Scale (SRS), "Patient motivation" (PM) and "Patient participation" (PP)
Notes	

Characteristics of ongoing studies *[ordered by study ID]*

Metz 2015

Trial name or title	Shared Decision Making in mental health care using Routine Outcome Monitoring as a source of information: a cluster randomised controlled trial
Methods	Multi-centre 2-arm cluster randomised controlled trial: pairs of teams from the same mental health organisation are randomly assigned to either the experimental or control conditions (matched pairs)
Participants	Clients attending specialised mental health care treated in subgroups by: age (adolescents, adults and elderly patients), diagnosis (psychotic, common mental and personality disorders); and setting (outpatient, day-clinic and clinic)
Interventions	Shared Decision Making (SDM) using Routine Outcome Monitoring (ROM) as a source of information (SDM-ROM model), using ROMs tailored to patient subgroup
Outcomes	Primary outcome: degree of decisional conflict, measured using Decisional Conflict Scale (DCS). Secondary outcomes: patient-clinician relationship assessed using Dutch version of Working Alliance Inventory Short Form (WAI-S); treatment outcome using either Manchester Short Quality of Live Measurement (MANSA-VN-16) for long-term patients or the Outcome Questionnaire (OQ-45) for short term patients

Starting date	August 2015
Contact information	Margot Metz, Trimbos Institute of Mental Health and Addiction, Utrecht, and GGZ Breburg, Tilburg, The Netherlands, email: m.metz@ggzbreburg.nl
Notes	

NCT01796223

Trial name or title	Effects of systematic patient feedback on therapy outcome and dropout: A randomized controlled study on adult out-patients at a community mental health centre
Methods	2 groups: Control (psychotherapy as usual); Intervention (psychotherapy along with feedback to therapist of a PROM administered at the beginning and end of every therapy session)
Participants	18 years or older Referred for treatment of mental disorder
Interventions	Partners for Change Outcome Management System (PCOMS) (KOR - Norwegian)
Outcomes	Health care utilisation (referral and drop out rate); symptom level; patient satisfaction; level of functioning; preferences for involvement in decision making; patient activation measure; use of health services (number of visits at General Practitioner or use of other health services)
Starting date	December 2012
Contact information	John Morten Koksvik, MD 0047 73 86 40 00 john.morten.koksvik@stolav.no Mariela Lara 0047 73 86 40 00 mariela.lara@stolav.no
Notes	Sponsors and Collaborators: Norwegian University of Science and Technology Principal Investigator: Mariela M Lara MA

NCT02023736

Trial name or title	Assessing psychotherapy outcome in treatment as usual versus treatment as usual with the STIC feedback system
Methods	2 groups: treatment-as-usual (TAU) versus TAU plus feedback to therapist through weekly online client questionnaires
Participants	Individuals, couples, and families
Interventions	Systemic Therapy Inventory of Change (STIC)
Outcomes	Change in mental health symptoms at termination, tailored to client demographics, including some or all of: Beck Depression Inventory II; Beck Anxiety Inventory; Outcome Questionnaire 45; Short-form 36 Health Survey; Revised Dyadic Adjustment Scale; Family Assessment Device; Strengths-Difficulties Questionnaire

NCT02023736 (*Continued*)

Starting date	December 2013
Contact information	Contact: Tara Latta:847 733 4300 ext 322 taralatta2008@u.northwestern.edu; Jacob Goldsmith, Ph.D. 847 733 4300 ext 860 jgoldsmith2@family-institute.org
Notes	

NCT02095457

Trial name or title	A randomised trial of routine computerised outcome and process clinical measures monitoring in mental health outpatient services: preparing for the planned public mental health reform in Israel
Methods	The suggested study is a 2-stage (implementation and intervention) open trial. 900 new outpatients in 'Shalvata' clinics will be recruited and randomised to intervention (ROM) and control groups. Assessment questionnaires will be filled periodically using 'CORE-NET', a computerised system enabling repeated measurements and feedback in a user-friendly and efficient manner
Participants	New patients attending clinic and beginning therapy
Interventions	Control group: Infrequently complete CORE-OM (once a year) and results not fed back to therapists Intervention group: Feedback of CORE-OM rating scale to therapists (completed between once a week and every 3 months)
Outcomes	Overall clinical wellbeing as measured by the CORE-OM rating scale Hospitalisation rates
Starting date	July 2014
Contact information	Ori Ganor, MD 972-54-5454886 origa1@clalit.org.il Lior Biran, Clinical Psychologist 972-54-4708886 liorbiran@gmail.com; liorbi@clalit.org.il
Notes	Sponsors and Collaborators: Shalvata Mental Health Center Principal Investigators: Ori Ganor MD, Lior Biran

NCT02656641

Trial name or title	Using the PHQ-9 and GAD-7 as feedback instruments in brief psychotherapy
Methods	Randomised controlled trial with 3 arms: Continuous Client Feedback (scores given to client and discussed with therapist); Continuous Self Feedback (scores given to to client only); Control (clients complete symptom and quality-of-life scales only before first session and before last or 10th session, whichever occurs first)
Participants	Aged 18-64, with diagnosis of major depressive disorder, generalized anxiety disorder, or adjustment disorder, undergoing brief psychotherapy
Interventions	Patient Health Questionnaire-9 (PHQ-9) and Generalized Anxiety Disorder 7-Item Scale (GAD-7) PROMs

NCT02656641 (*Continued*)

Outcomes	Primary outcome measures: Change in depressive symptoms on PHQ-9 and change in anxiety symptoms on GAD-7 at end of treatment (up to 10 weeks) Secondary outcome measures: WHO - Quality of Life BREF Scale for physical health, psychological health, social relationships, and environment; and Schwartz Outcome Scale
Starting date	November 2015
Contact information	Contact: Krystal G Ludwig, MA, 6104135983, kludwig@40christianacare.org David York, PhD, 3026230201, dyork@40christianacare.org, Christiana Healthcare, Wilmington, Delaware, United States
Notes	

NTR5466

Trial name or title	Routine Process Monitoring, systematic patient feedback in the primary and specialised mental healthcare
Methods	Randomised controlled trial with 2 arms: Routine Process Monitoring + Treatment as usual (TAU-RPM), and Treatment as Usual (TAU)
Participants	Patients aged 18 years or older, assigned to have psychological treatment in primary care or specialised mental healthcare
Interventions	Feedback of Session Rating Scale (SRS) and Outcome Rating Scale (ORS) scores completed in each treatment session
Outcomes	Primary outcomes: Outcome Questionaire 45 symptom score and Dutch Mental Health Continuum - Short Form (MHC-SF). Secondary outcomes: dropout, patient-satisfaction, duration of therapy, and treatment costs. Assessed at 5 weeks', 13 weeks', and 26 weeks' follow-up
Starting date	December 2015
Contact information	Dr AM Bovendeerd, Steenwijk, The Netherlands Tel: +31 (0)521 534140 email: b.bovendeerd@dimence.nl
Notes	

NTR5707

Trial name or title	Self-monitoring and personalised feedback as a tool to boost depression treatment
Methods	Randomised controlled trial with 3 arms: 'Do'-module (n = 50): patients report ESM data via their smartphone, 5 times a day for 28 days, with weekly feedback (to the patient) on positive affect (PA) and activities. 'Think'-module (n = 50): patients report ESM data via their smartphone, 5 times a day for 28 days, with weekly feedback (to the patient) on negative affect (NA) and thinking patterns. Control group (n = 50): patients on the wait list
Participants	Patients aged between 18 and 65 years for whom depression treatment is indicated by the practitioner

NTR5707 (*Continued*)

Interventions	Self-monitoring and personalised feedback through the Experience Sampling Method (ESM)
Outcomes	Change in depression symptom severity on self-report Inventory of Depressive Symptomatology; change in psychosocial functioning on Outcome Questionnaire 45; self-esteem and control over own lives on Dutch Empowerment questionnaire
Starting date	March 2016
Contact information	Dr JACJ Bastiaansen, Groningen, The Netherlands Tel: +31 (0)503 611169 email: j.bastiaansen@umcg.nl
Notes	

DATA AND ANALYSES

Comparison 1. Difference in outcome feeding back OQ-45 or ORS scores versus no feedback

Outcome or subgroup title	No. of studies	No. of participants	Statistical method	Effect size
1 Mean improvement in symptom scores: OQ-45 PROMS	9	3438	Mean Difference (IV, Random, 95% CI)	-1.14 [-3.15, 0.86]
2 Mean improvement in symptom scores: OQ-45 or ORS PROMs	12	3696	Std. Mean Difference (IV, Random, 95% CI)	-0.07 [-0.16, 0.01]
3 Number of treatment sessions received: all participants	7	2608	Mean Difference (IV, Random, 95% CI)	-0.02 [-0.42, 0.39]

Comparison 2. Subgroup analysis: Setting

Outcome or subgroup title	No. of studies	No. of participants	Statistical method	Effect size
1 Mean improvement in symptom scores by setting	12	3696	Std. Mean Difference (IV, Random, 95% CI)	-0.07 [-0.16, 0.01]
1.1 Multidisciplinary mental health care setting	7	1848	Std. Mean Difference (IV, Random, 95% CI)	-0.05 [-0.18, 0.07]
1.2 Psychological therapy setting	5	1848	Std. Mean Difference (IV, Random, 95% CI)	-0.10 [-0.23, 0.03]

Comparison 3. Subgroup analysis: Whether participants were given a formal diagnosis or not

Outcome or subgroup title	No. of studies	No. of participants	Statistical method	Effect size
1 Mean improvement in symptom scores by whether participants were given a formal diagnosis or not	12	3696	Std. Mean Difference (IV, Random, 95% CI)	-0.07 [-0.16, 0.01]
1.1 Mean improvement in symptom scores: participants given a formal diagnosis	3	1144	Std. Mean Difference (IV, Random, 95% CI)	-0.01 [-0.23, 0.21]
1.2 Mean difference in symptom scores: participants not given a formal diagnosis	9	2552	Std. Mean Difference (IV, Random, 95% CI)	-0.08 [-0.15, 0.00]

Comparison 4. Subgroup analysis: Feeback given to clinician, participant or both

Outcome or subgroup title	No. of studies	No. of participants	Statistical method	Effect size
1 Mean improvement in symptom scores: feedback given to clinician, participant or both	12	3696	Std. Mean Difference (IV, Random, 95% CI)	-0.07 [-0.16, 0.01]
1.1 Mean improvement in symptom scores: feedback given only to the clinician	2	140	Std. Mean Difference (IV, Random, 95% CI)	-0.17 [-0.63, 0.30]
1.2 Mean improvement in symptom scores: feedback given explicitly to both clinician and participant	6	862	Std. Mean Difference (IV, Random, 95% CI)	-0.12 [-0.30, 0.05]
1.3 Mean improvement in symptom scores: clinicians permitted or encouraged to share feedback with participant	6	2694	Std. Mean Difference (IV, Random, 95% CI)	-0.05 [-0.16, 0.06]

Comparison 5. Subgroup analysis: Whether feedback included treatment instructions or an algorithm

Outcome or subgroup title	No. of studies	No. of participants	Statistical method	Effect size
1 Mean improvement in symptom scores by whether feedback included treatment instructions or an algorithm	12	3696	Std. Mean Difference (IV, Random, 95% CI)	-0.07 [-0.16, 0.01]
1.1 Mean improvement in symptom scores: treatment instructions or algorithm	2	1184	Std. Mean Difference (IV, Random, 95% CI)	-0.03 [-0.14, 0.09]
1.2 Mean improvement in symptom scores: no treatment instructions or algorithm	10	2512	Std. Mean Difference (IV, Random, 95% CI)	-0.09 [-0.20, 0.02]

Comparison 6. Subgroup analysis: studies involving Michael Lambert versus studies not involving him

Outcome or subgroup title	No. of studies	No. of participants	Statistical method	Effect size
1 Mean improvement in symptom scores by whether studies involved Michael Lambert	9		Std. Mean Difference (IV, Random, 95% CI)	Subtotals only

1.1 Mean improvement in symptom scores: studies involving Michael Lambert	5	2032	Std. Mean Difference (IV, Random, 95% CI)	-0.06 [-0.15, 0.03]
1.2 Mean improvement in symptom scores: studies not involving Michael Lambert	4	1406	Std. Mean Difference (IV, Random, 95% CI)	-0.02 [-0.19, 0.15]

Comparison 7. Post hoc analyses - 'on track' and 'not on track' participants

Outcome or subgroup title	No. of studies	No. of participants	Statistical method	Effect size
1 Mean improvement in symptom scores: 'not on track' participants only	10	923	Std. Mean Difference (IV, Random, 95% CI)	-0.22 [-0.35, -0.09]
2 Number of treatment sessions received by 'on track' and 'not on track' participants	5	2114	Mean Difference (IV, Random, 95% CI)	0.06 [-0.91, 1.02]
2.1 Number of treatment sessions received by "on track" participants only	4	1633	Mean Difference (IV, Random, 95% CI)	-0.69 [-1.10, -0.29]
2.2 Number of treatment sessions received by "not on track" participants only	5	481	Mean Difference (IV, Random, 95% CI)	0.73 [-2.04, 3.50]

Analysis 1.1. Comparison 1 Difference in outcome feeding back OQ-45 or ORS scores versus no feedback, Outcome 1 Mean improvement in symptom scores: OQ-45 PROMS.

Review: Routine use of patient reported outcome measures (PROMs) for improving treatment of common mental health disorders in adults

Comparison: 1 Difference in outcome feeding back OQ-45 or ORS scores versus no feedback

Outcome: 1 Mean improvement in symptom scores: OQ-45 PROMS

Study or subgroup	Feedback N	Mean(SD)	No feedback N	Mean(SD)	Mean Difference IV,Random,95% CI	Weight	Mean Difference IV,Random,95% CI
Amble 2014	144	75.5 (28.6)	115	84.6 (25.1)		7.6 %	-9.10 [-15.65, -2.55]
De Jong 2012	206	67.43 (25.78)	207	67.53 (24.16)		12.0 %	-0.10 [-4.92, 4.72]
De Jong 2014	171	54.18 (23.99)	71	51.76 (26.84)		6.5 %	2.42 [-4.78, 9.62]
De Jong 2014	159	57.55 (25.91)	71	51.76 (26.84)		6.1 %	5.79 [-1.64, 13.22]
Hansson 2013	136	88.49 (17.82)	126	88.97 (18.53)		13.6 %	-0.48 [-4.89, 3.93]
Hawkins 2004	67	62.49 (25.82)	32	69.33 (23.42)		3.5 %	-6.84 [-17.04, 3.36]
Hawkins 2004	70	69.41 (24.56)	32	69.33 (23.42)		3.7 %	0.08 [-9.87, 10.03]
Lambert 2001	307	63.32 (23.83)	302	65.12 (22.31)		17.1 %	-1.80 [-5.47, 1.87]
Probst 2013	107	71.32 (25.66)	96	72.95 (25.27)		6.8 %	-1.63 [-8.64, 5.38]
Trudeau 2001	27	89.17 (22.43)	11	101.48 (27.32)		1.2 %	-12.31 [-30.54, 5.92]
Whipple 2003	499	58.15 (22.25)	482	58.56 (23.38)		22.0 %	-0.41 [-3.27, 2.45]
Total (95% CI)	**1893**		**1545**			**100.0 %**	**-1.14 [-3.15, 0.86]**

Heterogeneity: $Tau^2 = 2.63$; $Chi^2 = 13.30$, df = 10 (P = 0.21); $I^2 = 25\%$

Test for overall effect: Z = 1.12 (P = 0.26)

Test for subgroup differences: Not applicable

Analysis 1.2. Comparison 1 Difference in outcome feeding back OQ-45 or ORS scores versus no feedback, Outcome 2 Mean improvement in symptom scores: OQ-45 or ORS PROMs.

Review: Routine use of patient reported outcome measures (PROMs) for improving treatment of common mental health disorders in adults

Comparison: 1 Difference in outcome feeding back OQ-45 or ORS scores versus no feedback

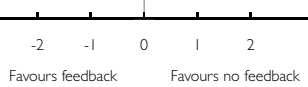
Outcome: 2 Mean improvement in symptom scores: OQ-45 or ORS PROMs

Study or subgroup	Feedback N	Mean(SD)	No feedback N	Mean(SD)	Std. Mean Difference IV,Random,95% CI	Weight	Std. Mean Difference IV,Random,95% CI
Amble 2014	144	75.5 (28.6)	115	84.6 (25.1)		8.4 %	-0.33 [-0.58, -0.09]
De Jong 2012	206	67.43 (25.78)	207	67.53 (24.16)		11.5 %	0.00 [-0.20, 0.19]
De Jong 2014	159	57.55 (25.91)	71	51.76 (26.84)		7.0 %	0.22 [-0.06, 0.50]
De Jong 2014	171	54.18 (23.99)	71	51.76 (26.84)		7.1 %	0.10 [-0.18, 0.37]
Hansson 2013	136	88.49 (17.82)	126	88.97 (18.53)		8.6 %	-0.03 [-0.27, 0.22]
Hawkins 2004	70	69.41 (24.56)	32	69.33 (23.42)		3.7 %	0.00 [-0.41, 0.42]
Hawkins 2004	67	62.49 (25.82)	32	69.33 (23.42)		3.6 %	-0.27 [-0.69, 0.15]
Lambert 2001	307	63.32 (23.83)	302	65.12 (22.31)		14.1 %	-0.08 [-0.24, 0.08]
Murphy 2012	59	-24.39 (7.13)	51	-23.77 (6.87)		4.4 %	-0.09 [-0.46, 0.29]
Probst 2013	107	71.32 (25.66)	96	72.95 (25.27)		7.2 %	-0.06 [-0.34, 0.21]
Reese 2009a	50	-31.28 (6.63)	24	-29.53 (7.26)		2.8 %	-0.25 [-0.74, 0.24]
Reese 2009b	45	-29.51 (9.58)	29	-24.33 (7.51)		2.9 %	-0.58 [-1.06, -0.10]
Trudeau 2001	27	89.17 (22.43)	11	101.48 (27.32)		1.4 %	-0.50 [-1.22, 0.21]
Whipple 2003	499	58.15 (22.25)	482	58.56 (23.38)		17.2 %	-0.02 [-0.14, 0.11]
Total (95% CI)	**2047**		**1649**			**100.0 %**	**-0.07 [-0.16, 0.01]**

Heterogeneity: $Tau^2 = 0.01$; $Chi^2 = 18.46$, df = 13 (P = 0.14); $I^2 = 30\%$

Test for overall effect: Z = 1.63 (P = 0.10)

Test for subgroup differences: Not applicable



Analysis 1.3. Comparison 1 Difference in outcome feeding back OQ-45 or ORS scores versus no feedback, Outcome 3 Number of treatment sessions received: all participants.

Review: Routine use of patient reported outcome measures (PROMs) for improving treatment of common mental health disorders in adults

Comparison: 1 Difference in outcome feeding back OQ-45 or ORS scores versus no feedback

Outcome: 3 Number of treatment sessions received: all participants

Study or subgroup	Feedback		No feedback		Mean Difference	Weight	Mean Difference
	N	Mean(SD)	N	Mean(SD)	IV,Random,95% CI		IV,Random,95% CI
Amble 2014	144	9.7 (8.6)	115	10.3 (9.2)		3.4 %	-0.60 [-2.79, 1.59]
De Jong 2014	140	36 (56.7)	63	33.5 (40.5)		0.1 %	2.50 [-11.22, 16.22]
De Jong 2014	144	27.5 (17.2)	63	33.5 (40.5)		0.2 %	-6.00 [-16.39, 4.39]
Hawkins 2004	70	8.2 (6.39)	32	8.66 (8.63)		1.5 %	-0.46 [-3.80, 2.88]
Hawkins 2004	67	7.79 (5.52)	32	8.66 (8.63)		1.5 %	-0.87 [-4.14, 2.40]
Lambert 2001	307	3.55 (3.89)	302	3.7 (3.87)		42.9 %	-0.15 [-0.77, 0.47]
Reese 2009a	50	6.27 (5.22)	24	5.99 (4.66)		2.9 %	0.28 [-2.08, 2.64]
Reese 2009b	45	8.02 (5.82)	29	5.79 (3.4)		3.7 %	2.23 [0.13, 4.33]
Whipple 2003	499	5.47 (4.86)	482	5.46 (4.88)		43.9 %	0.01 [-0.60, 0.62]
Total (95% CI)	**1466**		**1142**			**100.0 %**	**-0.02 [-0.42, 0.39]**

Heterogeneity: $Tau^2 = 0.0$; $Chi^2 = 6.64$, df = 8 (P = 0.58); $I^2 = 0.0\%$

Test for overall effect: Z = 0.08 (P = 0.93)

Test for subgroup differences: Not applicable

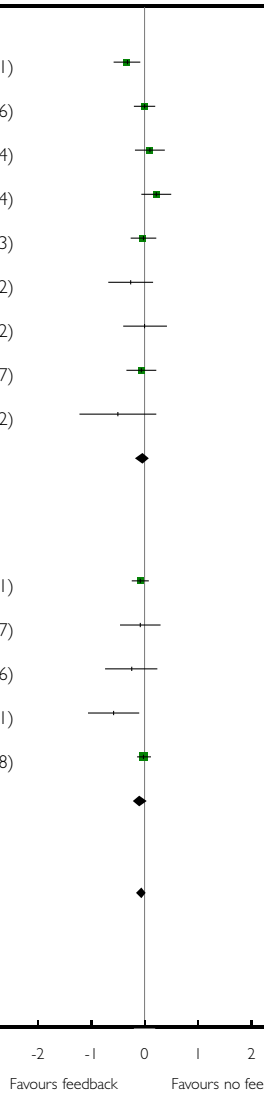
Analysis 2.1. Comparison 2 Subgroup analysis: Setting, Outcome 1 Mean improvement in symptom scores by setting.

Review: Routine use of patient reported outcome measures (PROMs) for improving treatment of common mental health disorders in adults

Comparison: 2 Subgroup analysis: Setting

Outcome: 1 Mean improvement in symptom scores by setting

Study or subgroup	Experimental N	Mean(SD)	Control N	Mean(SD)	Std. Mean Difference IV,Random,95% CI	Weight	Std. Mean Difference IV,Random,95% CI
1 Multidisciplinary mental health care setting							
Amble 2014	144	75.5 (28.6)	115	84.6 (25.1)		8.4 %	-0.33 [-0.58, -0.09]
De Jong 2012	206	67.43 (25.78)	207	67.53 (24.16)		11.5 %	0.00 [-0.20, 0.19]
De Jong 2014	171	54.18 (23.99)	71	51.76 (26.84)		7.1 %	0.10 [-0.18, 0.37]
De Jong 2014	159	57.55 (25.91)	71	51.76 (26.84)		7.0 %	0.22 [-0.06, 0.50]
Hansson 2013	136	88.49 (17.82)	126	88.97 (18.53)		8.6 %	-0.03 [-0.27, 0.22]
Hawkins 2004	67	62.49 (25.82)	32	69.33 (23.42)		3.6 %	-0.27 [-0.69, 0.15]
Hawkins 2004	70	69.41 (24.56)	32	69.33 (23.42)		3.7 %	0.00 [-0.41, 0.42]
Probst 2013	107	71.32 (25.66)	96	72.95 (25.27)		7.2 %	-0.06 [-0.34, 0.21]
Trudeau 2001	27	89.17 (22.43)	11	101.48 (27.32)		1.4 %	-0.50 [-1.22, 0.21]
Subtotal (95% CI)	**1087**		**761**			**58.5 %**	**-0.05 [-0.18, 0.07]**
Heterogeneity: Tau^2 = 0.01; Chi^2 = 12.70, df = 8 (P = 0.12); I^2 =37%							
Test for overall effect: Z = 0.84 (P = 0.40)							
2 Psychological therapy setting							
Lambert 2001	307	63.32 (23.83)	302	65.12 (22.31)		14.1 %	-0.08 [-0.24, 0.08]
Murphy 2012	59	-24.39 (7.13)	51	-23.77 (6.87)		4.4 %	-0.09 [-0.46, 0.29]
Reese 2009a	50	-31.28 (6.63)	24	-29.53 (7.26)		2.8 %	-0.25 [-0.74, 0.24]
Reese 2009b	45	-29.51 (9.58)	29	-24.33 (7.51)		2.9 %	-0.58 [-1.06, -0.10]
Whipple 2003	499	58.15 (22.25)	482	58.56 (23.38)		17.2 %	-0.02 [-0.14, 0.11]
Subtotal (95% CI)	**960**		**888**			**41.5 %**	**-0.10 [-0.23, 0.03]**
Heterogeneity: Tau^2 = 0.01; Chi^2 = 5.63, df = 4 (P = 0.23); I^2 =29%							
Test for overall effect: Z = 1.49 (P = 0.14)							
Total (95% CI)	**2047**		**1649**			**100.0 %**	**-0.07 [-0.16, 0.01]**
Heterogeneity: Tau^2 = 0.01; Chi^2 = 18.46, df = 13 (P = 0.14); I^2 =30%							
Test for overall effect: Z = 1.63 (P = 0.10)							
Test for subgroup differences: Chi^2 = 0.23, df = 1 (P = 0.63), I^2 =0.0%							



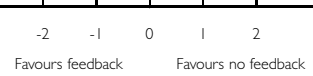
Analysis 3.1. Comparison 3 Subgroup analysis: Whether participants were given a formal diagnosis or not, Outcome 1 Mean improvement in symptom scores by whether participants were given a formal diagnosis or not.

Review: Routine use of patient reported outcome measures (PROMs) for improving treatment of common mental health disorders in adults

Comparison: 3 Subgroup analysis: Whether participants were given a formal diagnosis or not

Outcome: 1 Mean improvement in symptom scores by whether participants were given a formal diagnosis or not

Study or subgroup	Experimental N	Mean(SD)	Control N	Mean(SD)	Std. Mean Difference IV,Random,95% CI	Weight	Std. Mean Difference IV,Random,95% CI
1 Mean improvement in symptom scores: participants given a formal diagnosis							
Amble 2014	144	75.5 (28.6)	115	84.6 (25.1)		8.4 %	-0.33 [-0.58, -0.09]
De Jong 2012	206	67.43 (25.78)	207	67.53 (24.16)		11.5 %	0.00 [-0.20, 0.19]
De Jong 2014	159	57.55 (25.91)	71	51.76 (26.84)		7.0 %	0.22 [-0.06, 0.50]
De Jong 2014	171	54.18 (23.99)	71	51.76 (26.84)		7.1 %	0.10 [-0.18, 0.37]
Subtotal (95% CI)	**680**		**464**			**34.0 %**	**-0.01 [-0.23, 0.21]**
Heterogeneity: Tau^2 = 0.04; Chi^2 = 9.78, df = 3 (P = 0.02); I^2 =69%							
Test for overall effect: Z = 0.10 (P = 0.92)							
2 Mean difference in symptom scores: participants not given a formal diagnosis							
Hansson 2013	136	88.49 (17.82)	126	88.97 (18.53)		8.6 %	-0.03 [-0.27, 0.22]
Hawkins 2004	70	69.41 (24.56)	32	69.33 (23.42)		3.7 %	0.00 [-0.41, 0.42]
Hawkins 2004	67	62.49 (25.82)	32	69.33 (23.42)		3.6 %	-0.27 [-0.69, 0.15]
Lambert 2001	307	63.32 (23.83)	302	65.12 (22.31)		14.1 %	-0.08 [-0.24, 0.08]
Murphy 2012	59	-24.39 (7.13)	51	-23.77 (6.87)		4.4 %	-0.09 [-0.46, 0.29]
Probst 2013	107	71.32 (25.66)	96	72.95 (25.27)		7.2 %	-0.06 [-0.34, 0.21]
Reese 2009a	50	-31.28 (6.63)	24	-29.53 (7.26)		2.8 %	-0.25 [-0.74, 0.24]
Reese 2009b	45	-29.51 (9.58)	29	-24.33 (7.51)		2.9 %	-0.58 [-1.06, -0.10]
Trudeau 2001	27	89.17 (22.43)	11	101.48 (27.32)		1.4 %	-0.50 [-1.22, 0.21]
Whipple 2003	499	58.15 (22.25)	482	58.56 (23.38)		17.2 %	-0.02 [-0.14, 0.11]
Subtotal (95% CI)	**1367**		**1185**			**66.0 %**	**-0.08 [-0.15, 0.00]**
Heterogeneity: Tau^2 = 0.0; Chi^2 = 8.15, df = 9 (P = 0.52); I^2 =0.0%							
Test for overall effect: Z = 1.88 (P = 0.060)							
Total (95% CI)	**2047**		**1649**			**100.0 %**	**-0.07 [-0.16, 0.01]**
Heterogeneity: Tau^2 = 0.01; Chi^2 = 18.46, df = 13 (P = 0.14); I^2 =30%							
Test for overall effect: Z = 1.63 (P = 0.10)							
Test for subgroup differences: Chi^2 = 0.28, df = 1 (P = 0.60), I^2 =0.0%							

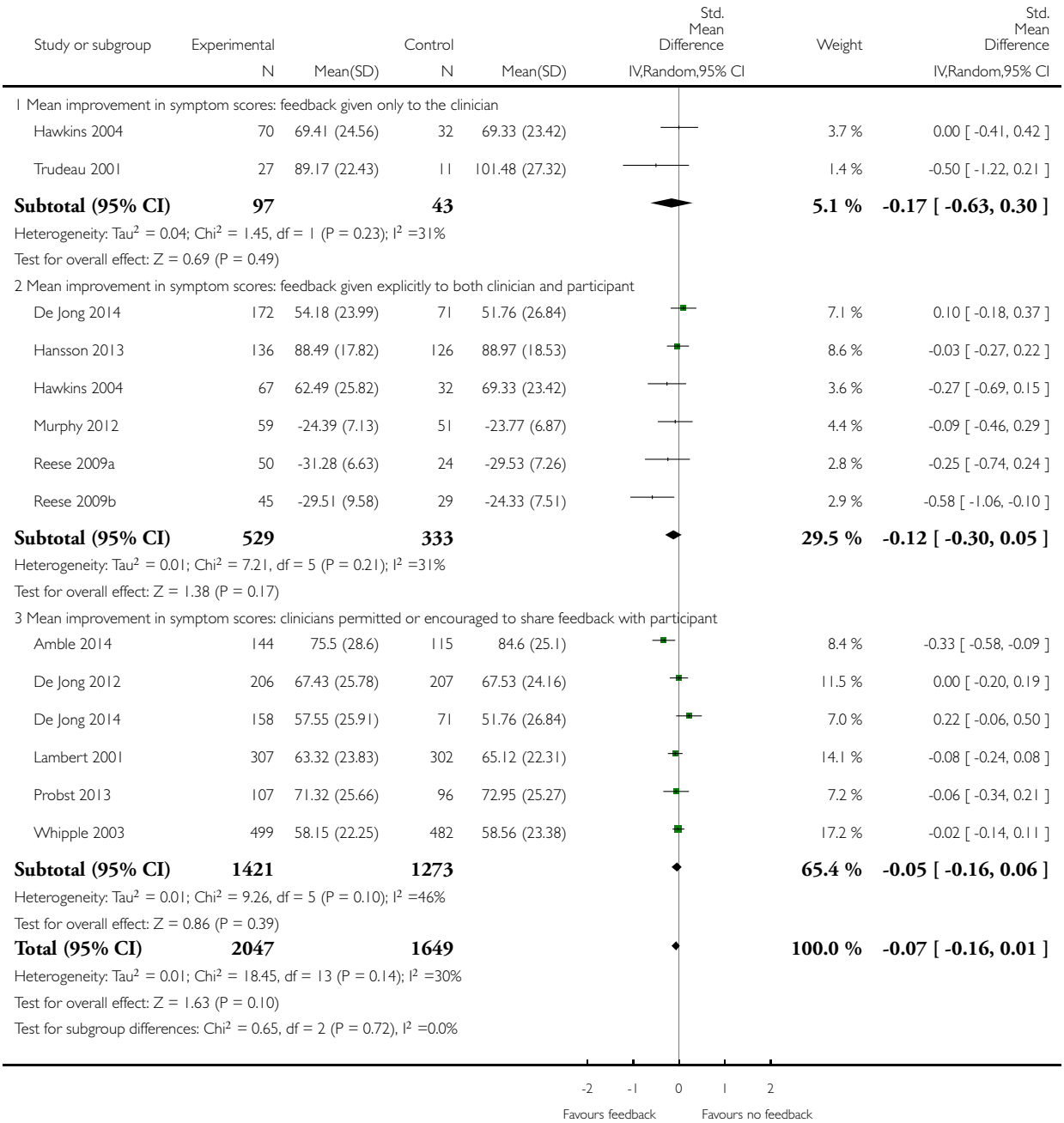


Analysis 4.1. Comparison 4 Subgroup analysis: Feeback given to clinician, participant or both, Outcome 1 Mean improvement in symptom scores: feedback given to clinician, participant or both.

Review: Routine use of patient reported outcome measures (PROMs) for improving treatment of common mental health disorders in adults

Comparison: 4 Subgroup analysis: Feeback given to clinician, participant or both

Outcome: 1 Mean improvement in symptom scores: feedback given to clinician, participant or both



Analysis 5.1. Comparison 5 Subgroup analysis: Whether feedback included treatment instructions or an algorithm, Outcome 1 Mean improvement in symptom scores by whether feedback included treatment instructions or an algorithm.

Review: Routine use of patient reported outcome measures (PROMs) for improving treatment of common mental health disorders in adults

Comparison: 5 Subgroup analysis: Whether feedback included treatment instructions or an algorithm

Outcome: 1 Mean improvement in symptom scores by whether feedback included treatment instructions or an algorithm

Study or subgroup	Experimental N	Mean(SD)	Control N	Mean(SD)	Std. Mean Difference IV,Random,95% CI	Weight	Std. Mean Difference IV,Random,95% CI
1 Mean improvement in symptom scores: treatment instructions or algorithm							
Probst 2013	107	71.32 (25.66)	96	72.95 (25.27)		7.2 %	-0.06 [-0.34, 0.21]
Whipple 2003	499	58.15 (22.25)	482	58.56 (23.38)		17.2 %	-0.02 [-0.14, 0.11]
Subtotal (95% CI)	**606**		**578**			**24.4 %**	**-0.03 [-0.14, 0.09]**
Heterogeneity: Tau² = 0.0; Chi² = 0.09, df = 1 (P = 0.77); I² =0.0%							
Test for overall effect: Z = 0.44 (P = 0.66)							
2 Mean improvement in symptom scores: no treatment instructions or algorithm							
Amble 2014	144	75.5 (28.6)	115	84.6 (25.1)		8.4 %	-0.33 [-0.58, -0.09]
De Jong 2012	206	67.43 (25.78)	207	67.53 (24.16)		11.5 %	0.00 [-0.20, 0.19]
De Jong 2014	172	54.18 (23.99)	71	51.76 (26.84)		7.1 %	0.10 [-0.18, 0.37]
De Jong 2014	158	57.55 (25.91)	71	51.76 (26.84)		7.0 %	0.22 [-0.06, 0.50]
Hansson 2013	136	88.49 (17.82)	126	88.97 (18.53)		8.6 %	-0.03 [-0.27, 0.22]
Hawkins 2004	70	69.41 (24.56)	32	69.33 (23.42)		3.7 %	0.00 [-0.41, 0.42]
Hawkins 2004	67	62.49 (25.82)	32	69.33 (23.42)		3.6 %	-0.27 [-0.69, 0.15]
Lambert 2001	307	63.32 (23.83)	302	65.12 (22.31)		14.1 %	-0.08 [-0.24, 0.08]
Murphy 2012	59	-24.39 (7.13)	51	-23.77 (6.87)		4.4 %	-0.09 [-0.46, 0.29]
Reese 2009a	50	-31.28 (6.63)	24	-29.53 (7.26)		2.8 %	-0.25 [-0.74, 0.24]
Reese 2009b	45	-29.51 (9.58)	29	-24.33 (7.51)		2.9 %	-0.58 [-1.06, -0.10]
Trudeau 2001	27	89.17 (22.43)	11	101.48 (27.32)		1.4 %	-0.50 [-1.22, 0.21]
Subtotal (95% CI)	**1441**		**1071**			**75.6 %**	**-0.09 [-0.20, 0.02]**
Heterogeneity: Tau² = 0.01; Chi² = 17.86, df = 11 (P = 0.08); I² =38%							
Test for overall effect: Z = 1.60 (P = 0.11)							
Total (95% CI)	**2047**		**1649**			**100.0 %**	**-0.07 [-0.16, 0.01]**

(*Continued* . . .)

(. . . *Continued*)

Study or subgroup	Experimental N	Mean(SD)	Control N	Mean(SD)	Std. Mean Difference IV,Random,95% CI	Weight	Std. Mean Difference IV,Random,95% CI
Heterogeneity: Tau² = 0.01; Chi² = 18.45, df = 13 (P = 0.14); I² =30%							
Test for overall effect: Z = 1.63 (P = 0.10)							
Test for subgroup differences: Chi² = 0.64, df = 1 (P = 0.42), I² =0.0%							

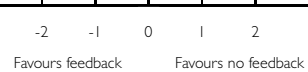
Analysis 6.1. Comparison 6 Subgroup analysis: studies involving Michael Lambert versus studies not involving him, Outcome 1 Mean improvement in symptom scores by whether studies involved Michael Lambert.

Review: Routine use of patient reported outcome measures (PROMs) for improving treatment of common mental health disorders in adults

Comparison: 6 Subgroup analysis: studies involving Michael Lambert versus studies not involving him

Outcome: 1 Mean improvement in symptom scores by whether studies involved Michael Lambert

Study or subgroup	Feedback N	Mean(SD)	No Feedback N	Mean(SD)	Std. Mean Difference IV,Random,95% CI	Weight	Std. Mean Difference IV,Random,95% CI
1 Mean improvement in symptom scores: studies involving Michael Lambert							
Hawkins 2004	70	69.41 (24.56)	32	69.33 (23.42)		4.4 %	0.00 [-0.41, 0.42]
Hawkins 2004	67	62.49 (25.82)	32	69.33 (23.42)		4.3 %	-0.27 [-0.69, 0.15]
Lambert 2001	307	63.32 (23.83)	302	65.12 (22.31)		30.5 %	-0.08 [-0.24, 0.08]
Probst 2013	107	71.32 (25.66)	96	72.95 (25.27)		10.1 %	-0.06 [-0.34, 0.21]
Trudeau 2001	27	89.17 (22.43)	11	101.48 (27.32)		1.5 %	-0.50 [-1.22, 0.21]
Whipple 2003	499	58.15 (22.25)	482	58.56 (23.38)		49.1 %	-0.02 [-0.14, 0.11]
Subtotal (95% CI)	**1077**		**955**			**100.0 %**	**-0.06 [-0.15, 0.03]**
Heterogeneity: Tau² = 0.0; Chi² = 3.02, df = 5 (P = 0.70); I² =0.0%							
Test for overall effect: Z = 1.30 (P = 0.19)							
2 Mean improvement in symptom scores: studies not involving Michael Lambert							
Amble 2014	144	75.5 (28.8)	115	84.6 (25.1)		20.0 %	-0.33 [-0.58, -0.09]



(*Continued* . . .)

(. . . *Continued*)

Study or subgroup	Feedback N	Mean(SD)	No Feedback N	Mean(SD)	Std. Mean Difference IV,Random,95% CI	Weight	Std. Mean Difference IV,Random,95% CI
De Jong 2012	206	67.43 (25.78)	207	67.53 (24.16)		23.9 %	0.00 [-0.20, 0.19]
De Jong 2014	158	57.55 (25.91)	71	51.76 (26.84)		17.8 %	0.22 [-0.06, 0.50]
De Jong 2014	172	54.18 (23.99)	71	51.76 (26.84)		18.1 %	0.10 [-0.18, 0.37]
Hansson 2013	136	88.49 (17.82)	126	88.97 (18.53)		20.3 %	-0.03 [-0.27, 0.22]
Subtotal (95% CI)	**816**		**590**			**100.0 %**	**-0.02 [-0.19, 0.15]**

Heterogeneity: $Tau^2 = 0.02$; $Chi^2 = 9.72$, df = 4 (P = 0.05); $I^2 = 59\%$

Test for overall effect: Z = 0.19 (P = 0.85)

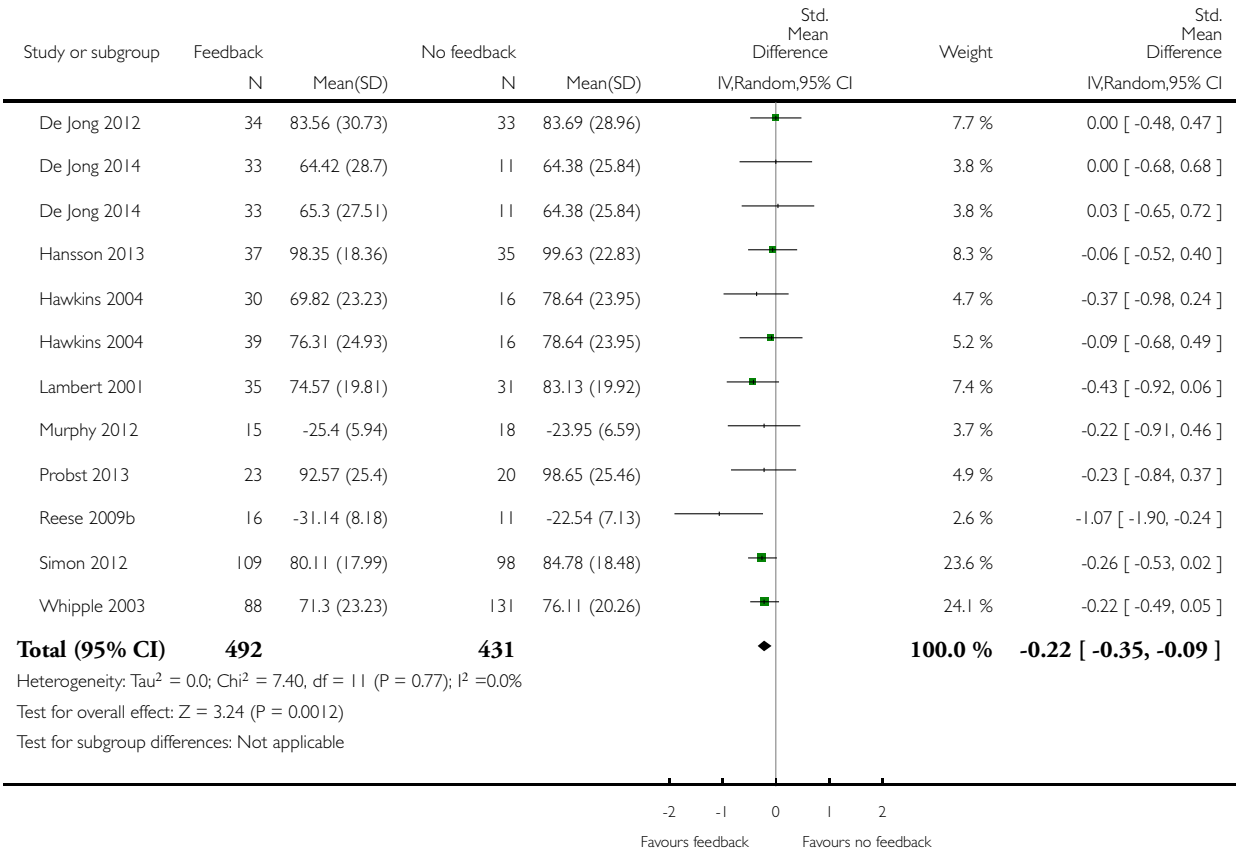
Test for subgroup differences: $Chi^2 = 0.18$, df = 1 (P = 0.67), $I^2 = 0.0\%$

Analysis 7.1. Comparison 7 Post hoc analyses - 'on track' and 'not on track' participants, Outcome 1 Mean improvement in symptom scores: 'not on track' participants only.

Review: Routine use of patient reported outcome measures (PROMs) for improving treatment of common mental health disorders in adults

Comparison: 7 Post hoc analyses - 'on track' and 'not on track' participants

Outcome: 1 Mean improvement in symptom scores: 'not on track' participants only

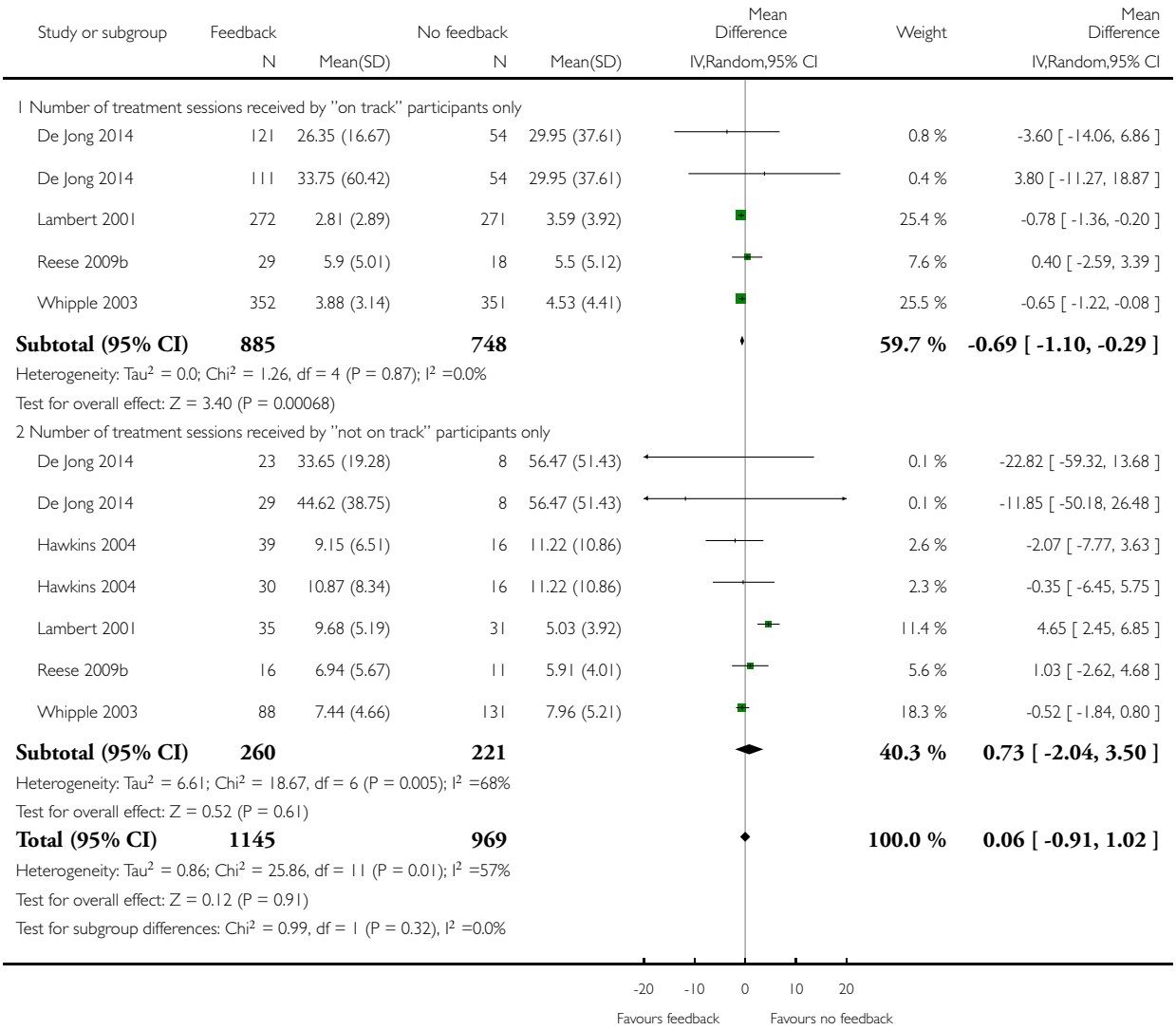


Analysis 7.2. Comparison 7 Post hoc analyses - 'on track' and 'not on track' participants, Outcome 2 Number of treatment sessions received by 'on track' and 'not on track' participants.

Review: Routine use of patient reported outcome measures (PROMs) for improving treatment of common mental health disorders in adults

Comparison: 7 Post hoc analyses - 'on track' and 'not on track' participants

Outcome: 2 Number of treatment sessions received by 'on track' and 'not on track' participants



ADDITIONAL TABLES

Table 1. Contact with investigators

Study ID	Response	Additional info	Outcome
Chang 2012	No	Failed to respond	Email sent to techang@partners.org and ayeung@partners.org on 22 January 2015 regarding randomisation process, CGI-S and PGI-S outcomes and criteria for diagnosis
De Jong 2012; De Jong 2014	Yes	Invited to become a co-author	Replied immediately to an email sent on 17 June 2015. Provided details of data on OQ-45 scores and further information on generation of random sequence and allocation concealment All data extraction was done by TK and MEG as KdJ became a co-author of the review
Haderlie 2012	Yes		Reply on 28 January 2015 in response to email sent on 22 January: "We collected the data in a naturalistic setting at 2 clinics with clients who were already in treatment in some cases. We did collect outcome data over the course of the study period, but we do not have first and last measurements in all cases. I also do not have specific information regarding the clients (such as diagnoses). Therapists were aware which condition they were in as the independent variable was whether or not they received progress feedback. Clients did not know which condition they were in"
Hansson 2013	Yes		Emails sent to helena.hansson@med.lu.se on 22 January 2015, 29 July 2015 and 10 August 2015 enquiring about separate outcome data per diagnostic group. Reply on 30 August with requested information
Mathias 1994	No	Failed to contact	Study too old - contact details of authors unobtainable
Probst 2013	No	Failed to respond	Email sent to thomas.probst@psychologie.uni-regensburg.de on 22 January 2015 regarding details of randomisation and blinding
Puschner 2009	Yes		Reply on 3 February 2015 in response to email sent 22 January 2015: query about breakdown of outcome data per diagnostic group, with tabulated data provided
Reese 2009a; Reese 2009b	Yes		Reply on 3 February 2015 in response to email sent 22 January 2015: to query about randomisation, allocation concealment and outcome blinding:

Table 1. Contact with investigators (*Continued*)

			Randomisation: "When a client was assigned to a therapist, the first client was randomised using a random number generator to either the feedback or TAU condition. The second client was then assigned to the other condition." Allocation concealment: "This was done by the person who assigned clients at the respective centres. This was done after enrolment into the study. Investigators and client participants could not foresee which condition a participant would be placed into." Outcome blinding: "The researchers did not know which condition participants were in until the time of analysis." Query over discrepancy in data presented in main text and table: "The table is correct and it should be 4.69. That is what was used in the analyses as well." Further data regarding number of treatment sessions and standard deviations provided on request in December 2015 Further information provided on diagnoses of study participants on 3 May 2016, confirming that more than 90% of them would have had qualifying clinical diagnoses of anxiety or depressive disorders, or both
Trudeau 2001	Yes		Reply on 26 March 2015 in response to email sent on 26 March 2015 enquiring about the details of managed care in place in some of the study participants, and about outcome data. "Managed care consisted of session limits and utilization review". Further email exchanges from 31 March to 2 April to enquire about blinding details. Reply on 13 May 2016 to email sent 13 May 2016, enquiring whether study participants would have met our review inclusion/exclusion characteristics, confirming that they would have met them
Lambert 2001; Simon 2012; Probst 2013; Whipple 2003	No	Failed to respond	Email sent to Michael_lambert@byu.edu on 22 January 2015 as listed as corresponding author on all four studies. Enquiries about randomisation procedure and allocation concealment. No reply regarding any of the studies was made. Further email sent to witold.simon@wp.pl and Michael Lambert on 17 July 2015, no reply received
Hawkins 2004	Yes		Email sent to eric.hawkins@va.gov on 18 November 2015, reply received: further data provided on OQ-

Table 1. Contact with investigators (*Continued*)

			45 outcomes and number of treatment sessions on 7 December 2015

APPENDICES

Appendix 1. PsycINFO search strategy (2014)

Ovid PsycINFO was initially searched using the following terms:
[Conditions]
1. *MENTAL DISORDERS/
2. (affective disorder* or common mental disorder* or mental health).ti,ab,id.
3. exp MAJOR DEPRESSION/
(major depression/ or anaclitic depression/ or dysthymic disorder/ or endogenous depression/ or postpartum depression/ or reactive depression/ or recurrent depression/ or treatment resistant depression/)
4. ATYPICAL DEPRESSION/
5. "DEPRESSION (emotion)"/
6. (depressi* or dysthymi*).ti,ab,id.
7. SEASONAL AFFECTIVE DISORDER/
8. exp ANXIETY/
(anxiety/ or computer anxiety/ or mathematics anxiety/ or performance anxiety/ or social anxiety/ or speech anxiety/ or test anxiety/)
9. exp ANXIETY DISORDERS/
(anxiety disorders/ or acute stress disorder/ or castration anxiety/ or death anxiety/ or generalized anxiety disorder/ or obsessive compulsive disorder/ or panic disorder/ or posttraumatic stress disorder/ or separation anxiety/)
10. exp NEUROSIS/
(neurosis/ or childhood neurosis/ or experimental neurosis/ or occupational neurosis/ or traumatic neurosis/)
11. OBSESSIONS/
12. PANIC ATTACK/ or PANIC/
13. (anxiety or compulsi* or obsess* or OCD or neurosis or neuroses or neurotic or panic or agoraphobi* or PTSD or posttrauma* or post-trauma* or acute stress or stress disorder* or psychotrauma* or psychological trauma*).ti,ab,id.
14. exp PHOBIAS/
15. (phobi* or fear or acrophobi* or arachnophobi* or claustrophobi*).ti,ab,id.
16. ADJUSTMENT DISORDERS/ or EMOTIONAL ADJUSTMENT/
17. adjustment reaction.ti,ab.id.
18. or/1-17
[Patient Reported Outcomes]
19. PROMS.ti,ab,id.
20. (patient reported outcome* or patient reported assessment* or patient reported symptom*).ti,ab,id.
21. (patient outcome*).ti,ab,id.
22. ((patient* or client* or tailored) adj2 feedback).ti,ab,id.
23. (patient* adj1 (self-assess* or self-report* or self-monitor*)).ti,ab,id.
24. (patient* adj2 progress*).ti,ab,id.
25. (client report*).ti,ab,id.
26. ((active or routine* or regular*) adj2 (feedback or measurement* or monitor*)).ti,ab,id.
27. (monitor* and feedback*).ti,ab,id.
28. (symptom* monitor* or treatment monitor*).ti,ab,id.

29. (monitor* adj2 (common mental disorder* or anxi* or compulsi* or obsess* or OCD or depressi* or neurosis or neuroses or neurotic or panic or phobi* or agoraphobi* or PTSD or posttrauma* or post-trauma* or acute stress or stress disorder* or trauma*)).ti,ab.
30. (follow up* assess*).ti,ab,id.
31. (needs assess*).ti,ab,id.
32. (outcome* adj3 (feedback or manag* or monitor*)).ti,ab,id.
33. severity questionnaire*.ti,ab,id,tm.
34. severity.ti,sh,tm. and (assess* or measure* or outcome* or questionnaire* or score*).ti.
35. (case management or enhanced care).ti,ab,id,sh.
36. "SEVERITY (disorders)"/
37. or/19-36
[RCT Filter]
38. Treatment Effectiveness Evaluation/
39. Clinical Trials/
40. Mental Health Program Evaluation/
41. Placebo/
42. placebo.ti,ab,id.
43. randomly.ab.
44. randomi#ed.ti,ab,id.
45. (control* adj3 (trial or study or group*1)).ti,ab,id.
46. factorial*.ti,ab.
47. allocat*.ti,ab.
48. assign*.ti,ab.
49. (crossover* or cross over*).ti,ab,id.
50. (quasi adj (experimental or random*)).ti,ab,id.
51. "2000".md.
52. (waitlist* or (wait* and list* and (control* or group))).ti,ab,id.
53. (treatment as usual or TAU or usual care or care as usual).ti,ab,id.
54. or/34-51
55. (18 and 37 and 54)
Key:
ab:abstract; id:key concepts; ti:title; tm:tests and measures
"2000".md.: methodology = treatment outcome/clinical trial

Appendix 2. Update searches (2015)

1. CCDANCTR-Studies and References Register was re-searched (*all years to 18 May 2015*) using the following terms:
#1. ((physician* or psychiatri* or psychotherapist* or therapist* or "primary care" or "general practi*") and (client* or patient* or oupatient*) near (feedback or feed-back)):ti,ab,kw,ky,mh,mc,emt
#2. (("psychotherapeutic outcome*" or "treatment outcome*") and (feedback or feed-back)):ti,ab,kw,ky,mh,mc,emt
#3. ("patient reported" near (information or outcome* or progress*)):ti,ab,kw,ky,mh,mc,emt
#4. (#1 or #2 or #3)
2. **The Cochrane Central Register of Controlled Trials (CENTRAL)** was searched (*all years to 18 May 2015*) using the following terms:
#1. ((psychiatri* or psychotherapist* or therapist*) and ((client* or patient* or oupatient*) near (feedback or feed-back))):ti,ab,kw
#2. ((psychiatri* or psychotherapist* or therapist*) and (patient-reported near (outcome* or progress))):ti,ab,kw
#3. (("psychotherapeutic outcome*" or "treatment outcome*") near (feedback or feed-back or (patient-reported near (information or outcome* or progress)))):ti,ab,kw
#4. ((physician or "primary care" or "general practi*") and ((client* or patient* or oupatient*) near (feedback or feed-back or progress))):ti,ab,kw
#5. MeSH descriptor: [MENTAL DISORDERS] explode all trees
#6. MeSH descriptor: [MENTAL HEALTH] explode all trees
#7. MeSH descriptor: [PSYCHOLOGICAL PHENOMENA and PROCESSES] explode all trees

#8. ((#1 or #2 or #3) or (#4 and (#5 or #6 or #7)))
OVID PsycINFO was re-searched (*all years to 18 May 2015*) using the following terms:
[Condition = 'receiving treatment/in therapy']
1. COUNSELING/
2. PSYCHOTHERAPY/
3. PSYCHOTHERAPEUTIC OUTCOMES/
4. TREATMENT OUTCOMES/
5. THERAPISTS/
6. "3310".cc. [*Classification Code:Psychotherapy & Psychotherapeutic Counselling*]
7. or/1-6
[Intervention]
8. FEEDBACK/
9. (feedback or feed-back).ti,id.10. or/8-9
[Condition + Intervention]
11. 7 and 10
12. ((physician* or psychiatri* or psychotherapist* or therapist* or primary care or general practi*) and ((client* or patient* or oupatient*) adj5 (feedback or feed-back))).ti,ab,id.
13. ((physician* or psychiatri* or psychotherapist* or therapist* or primary care or general practi*) and (patient reported adj3 (information or outcome*))).ti,ab,id.
14. (psychotherapeutic outcome* and (feedback or feed-back or (patient reported adj3 (information or outcome*)))).ti,ab,id.
15. or/11-14
[RCT filter]
16. TREATMENT EFFECTIVENESS EVALUATION/
17. CLINICAL TRIALS/
18. MENTAL HEALTH PROGRAM EVALUATION/
19. randomly.ab.
20. randomi#ed.ti,ab,id.
21. (control* adj3 (trial or study or group*1)).ti,ab,id.
22. "2000".md.
23. (waitlist* or (wait* and list* and (control* or group))).ti,ab,id.
24. (treatment as usual or TAU or usual care or care as usual).ti,ab,id.
25. or/16-24
[Condition + Intervention + RCTs]
26. 15 and 25
27. (3 or 4) and 10 *[no RCT filter]*
28. 26 or 27

F E E D B A C K

Feedback submitted, 15 July 2016

Summary

1) For many of the Cognitive Behaviour Therapy (CBT) programmes strongly supported in NICE recommendations, therapists are expected to use outcome and process measures every session in order to guide treatment.
For all the Cognitive Therapy (CT) for anxiety programmes that our group created (CT for panic, hypochondriasis, social anxiety disorder, PTSD) it is essential that therapists give process (inventories of condition specific automatic thoughts with belief ratings, attention & avoidance measures) as well as a relevant symptom measures before EVERY session in order to help decide on the content of the session. If a therapist fails to do this, they get heavily down rated on the therapist competency scales.

Improving Access to Psychological Therapies (IAPT) therapists are all supposed to give relevant symptom measures each session and encouraged to consider process measures. This is in line with the original implementation of CBT for depression as well as most of the most strongly evidence based anxiety programmes. In Beck's clinic (where I learned the depression treatment) patients have always been required to complete the Beck Depression Inventory and Beck Anxiety Inventory in the waiting room before the start of every session, along with other measures that the therapist might consider relevant. I would be gravely concerned that therapists are flying blind and insufficiently aware without this.

2) The few trials that look at giving therapists feedback from PROMs versus not giving feedback have always seemed strange to me as they have come from a way of working that seems completely alien to the competent delivery of the CBT programmes that have been evaluated in trials. The originators of those therapies mostly used process and outcome measures every session as an integral part of the clinical protocol. The feedback versus no feedback studies have often come from people with a counselling or psychodynamic background who haven't a tradition of using sessional measures to guide the content of interventions on a week-by-week basis.

3) The most common measure (OQ-45) is not one that we would consider useful for guiding CBT of a specific condition.

4) Feedback seems to be delayed, not like in IAPT where the patient fills in the questionnaire in the waiting room before a session and it is reviewed and acted on by the clinician immediately. Delayed/occasional feedback would be much less useful for guiding therapy.

5) Difficult to see how one could operate stepped care, a key aspect of IAPT, without recording outcomes. Perhaps for this reason, there don't seem to be any relevant RCTs looking at PROMS in such a complex, but routine clinical system.

6) The review doesn't discuss the value of standardised measures for identifying between and within service variation in outcomes and taking action to reduce them. We are working hard on this in IAPT, with some success.

Do you have any affiliation with or involvement in any organisation with a financial interest in the subject matter of your comment?
National Clinical Advisor for Improving Access to Psychological Therapies (IAPT) programme.

Reply

Thank you for your thoughtful and detailed feedback.
1) In our view, although inventories of condition specific thoughts, beliefs, attention, and avoidance measures may be an inseparable part of the therapeutic process in CBT, you could actually deliver it without routinely administering the set of symptom (PHQ-9, GAD-7) and social functioning (WSAS) questionnaires used at every IAPT session.
2) While it is true that a number of the studies included counselling or psychodynamic therapies, CBT was provided by at least some of the therapists in 11 of our included studies (Amble 2014, Berking 2006, de Jong 2012, de Jong 2014, Hawkins 2004, Lambert 2001, Murphy 2012, Reese 2009, Simon 2012, Trudeau 2001, and Whipple 2003). It is therefore possible to randomise patients to being monitored with PROMs even when using CBT. Furthermore, the large majority of patients with CMHDs don't get CBT - even within IAPT (which sees only 15% of all patients with CMHDs in the UK at most) only a proportion of those who are seen get full CBT - most get low intensity treatment. The large majority of all patients with CMHDs in the UK and other countries with well-developed primary care systems are prescribed antidepressants by their GPs, if they get any treatment at all. That is why we believe further research is needed in primary care as well as psychological therapy settings.
3) While the OQ-45 measure is non-specific for particular disorders - unlike the condition-specific scales used in IAPT - it has been used to monitor responses by therapists who used CBT in several of the included studies (Amble 2014, de Jong 2012, de Jong 2014, Hawkins 2004, Lambert 2001, Simon 2012, Trudeau 2001, and Whipple 2003).
4) In the included studies, the PROMs were usually completed immediately before, or during, the therapy session, so the scores would be available to guide therapy at that session, at least in psychological and mental health care settings. It is true that in the two primary care (family practice) studies included in our review (Mathias 1994 and Chang 2012), researchers working outside the practices administered the PROMs to participants, and fed back summary results to the clinicians, so there might have been some delay in giving feedback in those studies. We make the point in the review that systematising the administration of PROMs for CMHDs in primary care is more difficult, as only a minority of patients are attending for a CMHD, unlike in mental health or psychological therapy services, where all patients are attending for mental health problems and every patient can be given a PROM to complete as a routine on arrival.
5) Although the data collected in IAPT are used at least partly to determine whether patients should be stepped up to high intensity treatment, presumably they do not have to be completed at every therapy session, but could be done towards the end of a course of low

intensity treatment? It would be interesting, if it is possible to randomise patients to different frequencies of feedback within IAPT, to see whether completing PROMs at every session does confer an advantage, because the downside is that it may add to the burden of treatment for patients. Could a randomised trial of different frequencies of feedback with PROMs be conducted within the IAPT service?

6) We recognise that the data collected in IAPT are also used as performance measures to compare services with each other and to justify continued funding, but we were not looking at the use of PROMs to evaluate services in our review.

Contributors

Feedback submitted by: David Clark, University of Oxford
Response submitted by: Tony Kendrick, University of Southampton

WHAT'S NEW

Last assessed as up-to-date: 18 May 2015.

Date	Event	Description
5 August 2016	Feedback has been incorporated	Feedback submitted, 15 July 2016

CONTRIBUTIONS OF AUTHORS

TK had the idea for the study and led on the protocol. TK and RC linked with CCMD. TK co-ordinated the contributions from co-authors.

TK and MEG identified relevant abstracts and titles from searches, and selected studies for inclusion, involving MM in the discussion where there were differences.TK and MEG searched for and screened ongoing trials. TK, MEG, ALB and KDJ acquired the trial reports.

TK, MEG, MM, RC and SG assessed included studies for risk of bias. TK, MM and KDJ made contact with original study authors.

MEG, TK, LA, AB and ALB extracted data from study reports. MEG, TK and AG entered data into Review Manager. BS, AG and MEG re-checked the extracted data for accuracy against the study reports.

BS carried out the data analyses in discussion with TK and MEG. BS, TK, MEG and MM interpreted the results.

TK, MEG, BS and AG drafted the review. All the authors read and approved the final draft. TK is the guarantor of the review.

DECLARATIONS OF INTEREST

The authors have declared the following relevant interests.

Tony Kendrick: lead researcher on a grant of £244,892 from the National Institue for Health Research (NIHR) Research for Patient Benefit programme, for the PROMDEP feasibility study of using PROMs to monitor patients with depression in primary care.

Unpaid member of the NICE Indicator Advisory Committee for England, which considers evidence about the effectiveness of interventions such as the one subject to this Cochrane review, in recommending whether GPs and Clinical Commissioning Groups should be incentivised to implement them.

Magdy El-Gohary: none known.

Beth Stuart: co-applicant on the PROMDEP study.

Simon Gilbody: co-applicant on the PROMDEP study.

Rachel Churchill: leads and takes responsibility for the Cochrane Common Mental Disorders Group, which has supported parts of the review process and is funded by a grant from the National Institute of Health and Research (NIHR) in the UK.

Laura Aiken: none known.

Abhishek Bhattacharya: none known.

Amy Gimson: none known.

Anna Brütt: none known.

Kim de Jong: Two studies on which she was lead author are included in this review (De Jong 2012 and De Jong 2014), but TK and MEG carried out the data extraction from those studies.

Michael Moore: Co-applicant on the PROMDEP study.

SOURCES OF SUPPORT

Internal sources

- University of Southampton, UK.

Salary of Tony Kendrick and Michael Moore

- University of York, UK.

Salary of Simon Gilbody

- University of Bristol, UK.

Salary of Rachel Churchill

External sources

- NIHR School for Primary Care Research, UK.

Beth Stuart is supported by a post-doctoral award from the NIHR SPCR

- National Institute for Health Research, Not specified.

Magdy El-Gohary is supported by an NIHR GP In-Practice Fellowship

DIFFERENCES BETWEEN PROTOCOL AND REVIEW

In the protocol we referred to 'psychiatric practice' or 'specialised psychiatric practice' as one of the three service settings in which research has been done. We decided to revise the term to 'multidisciplinary mental health care' to reflect the fact that the clinicians treating patients with CMHDs in that setting were not only psychiatrists, but also included psychologists, mental health social workers, and mental health nurses.

A decision was made to include studies in which the diagnoses of the large majority of participants were reported as CMHDs, even if a proportion of participants were not given a specific diagnosis, or were reported as having 'relationship' or 'interpersonal' difficulties, 'somatoform disorders', 'other' diagnoses not further specified, or 'administrative codes'. This was a change from the protocol as we planned originally to include only studies with participants diagnosed with one of the disorders listed in the protocol, but after discussion within the review group we decided to include these studies as they included a large majority of participants with CMHDs, and we considered that excluding them would be to the detriment of the review, through the omission of data that would otherwise be valid and available.

A decision was also made to include studies where a small number of participants carried a diagnosis of exclusion including substance misuse, eating disorder or psychosis, where the total number of such participants constituted less than 10% of the total study population. This was because many of the larger studies included a small number of these participants, and again we considered that excluding these studies would be to the detriment of the review through omission of data that would otherwise be valid and available.

Where studies did not report the diagnoses of participants, we decided to contact the corresponding authors to request any available information on the participants' diagnoses, and whether they would have met the review inclusion and exclusion criteria. This was an addition to the protocol agreed after discussion within the review study group, again in order not to omit data that would otherwise be valid and available.

Consequently, we agreed to carry out additional, post-hoc sensitivity analyses of the meta-analyses of studies using the OQ-45 or ORS, omitting three studies included in the meta-analyses which did not report the specific diagnoses of their participants (Reese 2009a; Reese 2009b; Trudeau 2001), and five which did not assign a specific diagnosis to 20% or more of their participants (Amble 2014; De Jong 2014; Lambert 2001; Murphy 2012; Trudeau 2001; Whipple 2003).

In the protocol for the review, we planned to use logistic regression to calculate and present odds ratios (ORs) and associated 95% CIs for dichotomous (binary) outcomes, such as changes in antidepressant drug prescriptions or referrals for psychological or psychiatric treatment. However, in the event, dichotomous outcomes were uncommon findings in the review and therefore it was agreed such data should be presented in narrative form only.

We decided post-hoc to conduct an additional subgroup analysis, comparing studies involving Michael Lambert, the originator and owner of the OQ-45 system, with studies not involving him, to explore whether potential benefits of the system were identified in independent evaluations. This was because the OQ-45 was the PROM used in the large majority of studies in the meta-analyses, and Michael Lambert was author or co-author of a significant proportion of those studies.

Two further post-hoc subgroup comparisons were also agreed between the authors of subgroups of participants who were identified in a number of studies as being at higher or lower risk for treatment failure, which was determined by the trajectory of their initial response to therapy. The low risk group was described as 'on-track' (OT) for a good clinical response, and the high risk group as 'not on track' (NOT) or alternatively 'at risk', 'signal cases', or 'signal alert cases'. This was because several studies using the OQ-45 or ORS PROM systems reported positive findings among the NOT participants in the absence of significant findings among their samples overall. One comparison included only the NOT subgroup, comparing outcomes in terms of symptom scores between feedback and non-feedback arms. The second comparison included both the OT and NOT subgroups, comparing the number of treatment sessions received between feedback and non-feedback arms, and including a formal test for subgroup differences to look for evidence of differences between OT and NOT subgroups. This was also a change from the protocol, as the number of treatment sessions was a secondary outcome, and originally we planned to conduct subgroup analyses restricted to the primary outcomes (namely symptoms and adverse effects).

INDEX TERMS

Medical Subject Headings (MeSH)

*Patient Outcome Assessment; Antipsychotic Agents [therapeutic use]; Feedback; Mental Disorders [psychology; *therapy]; Quality of Life; Randomized Controlled Trials as Topic; Suicidal Ideation

MeSH check words

Adult; Aged; Female; Humans; Male; Middle Aged