Disagreement between self-administered depression questionnaires and patients' own views of their recovery: a cohort study

Background: Self-administered questionnaires are widely used in primary care and other clinical settings to assess the severity of depressive symptoms and monitor treatment outcomes. Qualitative studies have found that changes in questionnaire scores might not fully capture patients' experience of changes in their mood but there are no quantitative studies of this issue.

Aims: We examined the extent to which changes in scores from depression questionnaires disagreed with primary care patients' perceptions of changes in their mood and investigated factors influencing this relationship.

Methods: Prospective cohort study assessing patients on four occasions, two weeks apart. Patients (N=554) were recruited from primary care surgeries in three UK sites (Bristol, Liverpool and York) and had reported depressive symptoms or low mood in the past year (68% female, mean age 48.3 (SD 12.6)). Main outcome measures were changes in scores on Patient Health Questionnaire (PHQ-9) and Beck Depression Inventory (BDI-II) and the patients' own ratings of change,

Results: There was marked disagreement between clinically important changes in questionnaire scores and patient-rated change, with disagreement of 51% (95% CI 46% to 55%) on PHQ-9 and 55% (95% CI 51% to 60%) on BDI-II. Patients with more severe anxiety were less likely, and those with better mental and physical health related quality of life more likely, to report feeling better, having controlled for depression scores.

Conclusion: Our results illustrate the limitations of self-reported depression scales to assess clinical change. Clinicians should be cautious in interpreting changes in questionnaire scores without further clinical assessment.

Keywords: depression, primary care, PHQ-9, BDI-II, cohort.

Introduction

Self-administered screening questionnaires that assess the severity of depressive symptoms have been recommended in UK primary care and in North America and some parts of Europe[59,60]. These recommendations were made in response to concerns that depression is under-diagnosed and under-treated in primary care, with the aim of improving detection and monitoring treatment response. In 2006 the Quality Outcomes Framework (QOF) in the UK incentivised the use of three questionnaires: the Patient Health Questionnaire (PHQ-9), the Beck Depression Inventory (BDI-II) and the Hospital Anxiety and Depression Scale (HADS). These questionnaires are no longer incentivised but remain widely used in UK primary care and continue to influence treatment decisions [59]. The PHQ-9 along with other questionnaires is also used as a routine outcome measure in Improving Access to Psychological Therapies (IAPT) services in the UK [61].

Self-administered depression questionnaires have been compared to diagnostic assessments and their sensitivity and specificity is fairly good, at around 80% [62,63]. However, their use in clinical settings has been criticized [64,65]. One concern is that changes in scores might not fully capture the patient's experience of improvement or deterioration in their mood. Such disagreement has important implications for treatment decisions and patient-centred care [66,67].

Clinicians routinely ask patients whether their condition has improved, deteriorated or stayed the same [68,69]. Patient-rated change is measured in research settings with a single-item question, which asks patients retrospectively about how their whole condition has changed compared to a previous occasion, rather than asking about individual symptoms [68,69].

We have conducted qualitative studies of people whose self-rated changes in mood differed from their responses to self-administered depression scales [66,67]. Patients explained the disagreement as resulting from the presence of co-morbid conditions, negative and positive life events, changes in social support and changes in quality of life [66]. This supports other qualitative findings that patients often state that scales such as the PHQ-9 do not fully capture their experience of illness [67]. We are not aware of any similar qualitative or quantitative investigations of this question.

In this study we used a cohort of patients recruited from primary care to investigate the extent to which responses to the PHQ-9 and Beck Depression Inventory (BDI-II) disagreed with patients' perceptions of changes in their mood, assessed using a patient-rated change scale. We also investigated factors that might influence patient reports of self-improvement having controlled for their responses on the PHQ-9 and BDI-II.

Methods

Participants

Participants were recruited from General Practice (GP) surgeries in three UK sites: Bristol, Liverpool and York. Computerised records were used to identify patients aged 18 to 70 who

had reported low mood, depressive episodes, depressed mood, depressive symptoms or a major depressive episode in the past year, irrespective of any treatment. We excluded patients who: were diagnosed with bipolar disorder, psychosis or eating disorder; had alcohol or substance use problems; were unable to complete study questionnaires; or were 30 weeks or more pregnant. 7,721 patients were sent an information letter and 1,470 (19%) replied. Of these, 821 were willing to be contacted, 23 (3%) of whom were ineligible. The remaining 798 were contacted to arrange an interview. Of these, 563 consented (38%) and 559 (38%) were interviewed (4 could not be contacted). Data were collected at four time-points, two weeks apart (baseline and follow-up 1, 2 and 3). Patients and public representatives were involved in management and steering groups for the PANDA programme grant and gave input into the design, conduct and interpretation of the study.

Ethical Approval

All participants provided written informed consent and ethical approval was obtained from NRES Committee South West - Central Bristol. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975 as revised in 2008.

Measures

Depressive symptoms: The Patient Health Questionnaire (PHQ-9) and Beck Depression Inventory-II (BDI-II) were completed at each time-point. The PHQ-9 is a 9-item self-administered measure of depressive symptoms in the past two weeks and scores range from 0-27 [70]. Internal consistency was high at each time-point (Cronbach's alpha 0.89 to 0.92).

The BDI-II is a 21-item self-administered measure of the severity of depressive symptoms in the past two weeks [63] and scores range from 0 to 63. Internal consistency was high at each time-point (Cronbach's alpha 0.93 to 0.95). Higher scores indicate more severe depressive symptoms.

Patient-rated change: We used a single-item question based on 'Global Rating Scales' that are routinely used in musculoskeletal and chronic pain research and have high reliability and validity [68,69]. Participants were asked 'compared to when we last saw you 2 weeks ago how have your moods and feelings changed?' Response options were: 'I feel a lot better' (1), 'I feel slightly better' (2), 'I feel about the same' (3), 'I feel slightly worse' (4), 'I feel a lot worse' (5). We used 'moods and feelings' instead of 'depression' because many people might not consider themselves "depressed" and this wording should encourage a more general response. Our qualitative studies found evidence that patients viewed this question as more open-ended and explorative, stating that it allowed them to 'sum up' their mental health and express themselves outside of the parameters of the questionnaires [66]. The patient-rated change scale was completed twice at each time-point, at the beginning and end of the questionnaire. Test-retest reliability was good with kappa (quadratic weights) of 0.89. The scale, or similar, has been used in prior Randomised Controlled Trials [66,67,71].

Anxiety: The Generalised Anxiety Disorder Assessment (GAD-7) [72] was completed at each time-point and is a 7-item self-administered measure of the severity of anxiety symptoms in the past two weeks, scores ranging from 0 to 21. Higher scores indicate more severe symptoms.

Physical and Mental Health-Related Quality of Life: The 12-item Short-Form Health Survey (SF-12) [73] was administered at each time-point. Separate physical and mental health-related

quality of life scores were derived [73]. Scores range from 0 to 100, higher scores indicating better quality of life.

Negative Life Events: At baseline (only), participants were asked, using a self-administered computerized questionnaire, whether they had experienced the following in the previous 6 months: (i) bereavement, (ii) separation or divorce, (iii) a serious illness or injury, (iv) victimisation (mugging, burglary, serious assault), (v) being in trouble with the law, (vi) debt, (vii) a serious dispute with a family member or friend, or (viii) being made compulsorily redundant from work. Due to the low frequency, a binary variable was created (none or 1 or more).

Social Support: At baseline (only), participants completed eight questions as part of the self-administered computerized questionnaire relating to: (i) feeling loved, (ii) having others that can be relied on, (iii) feeling accepted, (iv) feeling supported, (v) having others to talk to, (vi) having others that make them happy, (vii) having others that care what happens to them, and (viii) having others that make them feel an important part of their lives. Each question used a three-point scale (1) not true, (2) partly true, and (3) certainly true. Scores were summed and ranged from 1 to 24, higher scores indicating more social support.

Potential confounders: We adjusted for variables previously shown to be associated with depressive symptoms, and site. Demographic variables (age, sex, ethnicity, employment status, financial status, and education level) were measured at baseline. Due to small numbers, ethnic minority status was a binary variable. Employment status was categorised as employed, unemployed not by choice, and unemployed by choice. Financial status was three categories: low (‘Finding it very difficult to make ends meet’ and ‘Finding it difficult to make

ends meet'), medium ('Just about getting by') and high ('Living comfortably' and 'Doing alright'). Education level was seven categories, from no qualifications to higher degree.

Statistical Analyses

Identifying disagreement between questionnaire scores and patient-rated change

To calculate change scores, mean PHQ-9 and BDI-II scores at each follow-up time-point were subtracted from mean scores at the previous time-point (to correspond to the patient-rated change scale which asks about change over the last two weeks). Possible change scores ranged from -27 to +27 for PHQ-9 and -63 to +63 for BDI-II. Greater negative scores indicated improvement and greater positive scores indicated deterioration.

We used the Minimal Clinically Important Difference (MCID), the smallest change in symptoms meaningful to patients, to assess extent of disagreement [71]. The MCID has been estimated in the PANDA cohort to be around a 20% reduction in PHQ-9 or BDI-II scores (manuscript in preparation). We used the MCID of a 20% reduction or increase in questionnaire scores to create the following categories: clinically important decreases (a decline in scores of 20% or more), no clinically important change (a decline or increase in scores smaller than 20%), and clinically important increases (an increase in scores greater than or equal to 20%) [71]. For each response option on the patient-rated change scale, we report the proportion of patients in each of the above MCID categories.

We defined disagreement as (i) a clinically important change in PHQ-9/ BDI-II scores and a rating of change response that indicated either no change or a change in the opposite direction

(ii) no clinically important change in PHQ-9/ BDI-II scores and a rating of change response that indicated a change in either direction. The proportion of patients showing some form of disagreement overall was calculated overall by dividing the total number of people showing disagreement by the total number of people. Proportion disagreement was also calculated within each patient-rated response category. Quadratic weighted and unweighted kappa values were used to test agreement between patient rating of change responses and MCID categories. In a prior manuscript we had identified a MCID of 15% for the BDI-II [71] so we conducted sensitivity analyses with this estimate.

Reliability of disagreement

We further examined the extent of disagreement by tabulating the proportion of participants scoring within each category of the patient-rated change scale with the equivalent proportion scoring a corresponding change on the PHQ-9/BDI-II (supplementary analyses). For example if 10% of patients reported feeling much better, this was tabulated against the top 10% of change scores on the PHQ9/BDI-II and so on for the percentage who reported feeling slightly better, the same, slightly worse or worse. Quadratic weighted and unweighted kappa values were used to test agreement between these proportions.

Variables that influence disagreement

We used a binary outcome (feeling better versus same or worse) to reflect that neither feeling the same nor worse is a good clinical outcome. As the patient-rated change scale asks about the last two weeks, we could construct logistic models with the 2, 4 or 6 week follow-up as the outcome. We adjusted for binary clinically important change (20% change in scores or not) over the previous two-weeks. This binary variable reduced collinearity between

depression scores and other exposures (e.g. anxiety) and was consistent with our approach to clinically important change and disagreement.

For exposures measured at multiple time-points (anxiety, mental and physical-health related quality of life,) we did a principal components analyses of the exposure at the current and preceding time-points. Principal components analysis (PCA) can be used to transform two correlated variables into orthogonal (uncorrelated) factors or 'principal components.' The first component is a function of the average score on each variable. The second component is uncorrelated with the first, and is a function of the difference between two scores [74]. Models were adjusted for confounders known to be associated with depressive symptoms (age, sex, ethnicity, education level, current use of antidepressants and marital, financial and employment status) and site. All analyses were conducted using STATA 14.

Role of the funding source

The funding source had no role in study design, data collection, data analysis, interpretation or writing of the report. The corresponding author had full access to all data used in the study, and final responsibility for the decision to submit for publication.

Results

Due to extensive missing data at baseline 5 patients were excluded, leaving 554 for analyses. At follow-ups one, two, and three: 476 (86%); 443 (80%), and 430 (78%) provided data respectively. Baseline sample characteristics are presented in Table 1. Patients were aged 18 to 71 (mean 48·30, SD 12·56), 68% female and 96% white.

Identifying disagreement between questionnaire scores and patient-rated change

Disagreement between questionnaire scores and the patient-rated change scale was similar across time-points, so data from baseline to follow-up 1 are presented for brevity.

Depression change scores according to patient-rated change

Change in depression questionnaire scores were related to patients' responses on the rating scale. Patients who reported 'feeling a lot better' had the largest mean decrease in scores, and patients who reported 'feeling a lot worse' the largest increase (Table 2, first row in PHQ9 and BDI-II sections).

Clinically important change in depression scores according to patient-rated change

When clinically important differences in depression scores were compared to patient ratings, there was evidence of disagreement. The proportion of patients showing each type of clinically important change in questionnaire scores (increase, no change, decrease), in comparison to their responses is presented in Table 2.

Disagreement was most common in patients who reported feeling worse on the patient-rated change scale. PHQ-9 scores showed no change or an improvement for 76% (95% CI: 66% to 83%) of those who reported 'feeling slightly worse', and 81% (95% CI: 54% to 94%) of those who reported 'feeling a lot worse' (Table 2, last row in PHQ-9 section). These results were very similar for the BDI-II (Table 2, last row in BDI-II section). Disagreement was also common in patients who reported feeling better. PHQ-9 scores remained the same or deteriorated in 65% (95% CI: 55% to 74%) of those who reported 'feeling slightly better', and 53% (95% CI: 37% to 67%) of those who reported 'feeling a lot better' (Table 2, last row

in PHQ-9 section). Disagreement was lower for patients who reported feeling better on the BDI-II: 43% (95% CI: 34% to 53%) for those reporting feeling slightly better and 28% (95% CI: 16% to 43%) for those reporting feeling much better (Table 2, last row in BDI-II section). Overall, the proportion of people showing some form of disagreement was 51% (95% CI: 46% to 55%) on the PHQ-9 and 55% (95% CI: 51% to 60%) on the BDI-II. When using a more stringent minimal clinically important difference of 15%, results were comparable (Supplementary Table 1).

Quadratic weighted Kappa scores indicated agreement between patient ratings and the categories generated from the change scores ranging from 81·2-83·6% for the PHQ-9 and 78·6-83·1% the BDI-II. Unweighted Kappa scores indicated low levels of agreement (3·9-7·6%) for PHQ-9 and BDI-II.

Reliability of disagreement

Results were similar when the proportion of patients scoring within each category of the patient-rated change scale were compared with the relative proportion of patients scoring within these ranges on the PHQ-9 and BDI-II (Supplementary Table 2). High agreement was observed between the patient-rated change scale and PHQ-9/BDI-II, with weighted kappa values indicating agreement ranging from to 91·4% to 93·1% across time-points. Unweighted kappa values indicated poorer agreement (37·9-42·4%). We found no evidence that disagreement differed according to gender (results available on request).

Variables that influence disagreement

Results for the PHQ-9 are shown in Table 3 and for the BDI-II, Table 4. We found evidence that an increase in anxiety symptoms was associated with a decreased odds of reporting

feeling better after controlling for changes in depressive symptoms. This was consistent across time-points, for PHQ-9 and BDI-II. For example at follow-up 1, a four-point increase in anxiety scores was associated with a 0·67 (95% CI 0·55 to 0·82) decrease in the odds of feeling better, having controlled for change in PHQ-9 scores.

We also found consistent evidence that improved mental and physical health related quality of life was associated with increased odds of reporting feeling better after controlling for changes in depressive symptoms. For example at follow-up 1, an eight-point increase in mental health related quality of life was associated with a 1·43 (95% CI 1.11 to 1.61) increase in the odds of feeling better. For physical health related quality life this odds ratio was 1.28 (1.08 to 1.54). There was no evidence of an influence of negative life events or social support on the likelihood of reporting improvement (Tables 3 and 4). We found no evidence that any of these associations differed according to gender (available on request).

Discussion

Summary of findings

We found evidence that changes in scores on self-administered depression questionnaires often differ from patients' own views of changes in their mood. Over 50% of people evidenced some form of disagreement between their questionnaire scores and self-rated mood. Even though, on average, there is fairly good agreement between change in depressive

symptoms and self-rated changes in mood, our results suggest that applying these questionnaires to individual patients will be prone to error.

Patients with more severe anxiety symptoms were less likely, and those with better mental and physical health related quality of life more likely, to report feeling better having controlled for their depression questionnaire scores. Our results support the idea that self-administered scales only capture a subset of the subjective experience that contributes to patient-rated change and suggests that relying solely upon responses to self-administered scales could be misleading in a large proportion of situations.

Strengths and Limitations

We set broad and inclusive entry criteria to reflect the patients consulting for depression in primary care. The Minimal Clinically Important Difference (MCID) allowed us to infer that differences were clinically important, though we acknowledge that the MCID is itself an average determined by reference to patients self-rated change. Our results indicate that such average MCIDs are difficult to apply in individual cases, even if they are valuable overall in planning and interpreting studies.

The depression questionnaires and patient-rated change scale will be subject to measurement error, which could be a potential source of disagreement. Multi-item scales with specific prompts might be more reliable [68], but the reliability of the patient-rated change scale was good. There could be other reasons for disagreement. The patient-rated change scale asks retrospectively about change and recall might be poor [75]. However, the recall period (2 weeks) was the same for the depression questionnaires and patient-rated change scale. 'Response shift' is the concept that answers will differ across time not because the condition

has changed but because the opinion on what the condition means has changed [76]. This might also lead to disagreement, if it occurred. Finally, it is unclear which aspects of the patients' condition have informed response to the patient-rated change scale. However, these points are largely concerned with explaining differences between the two contrasting approaches to assessment rather than casting doubt on our conclusions.

There was a low response rate for the study and this might have affected the representativeness of our target population which was patients seeking help in primary care. However, it seems unlikely that our method of recruitment and the low response rate would inflate the level of disagreement although we cannot rule out that possibility. Our sample was from the UK and predominantly white and this may limit generalizability. Finally, there was attrition though retention was good with 78% at the final follow-up.

These quantitative findings are partly consistent with our previous qualitative findings [66,67]. Of course, the PHQ-9 and BDI-II only measure depression symptoms so it is unsurprising that anxiety should affect patient-rated change in mood and feelings independently. Given the co-occurrence of depression and anxiety it is important to recognize that, from the patients' perspective, changes in anxiety will also be important.

The PHQ-9 and BDI-II are recommended for assessment of depressive illness and treatment response in UK primary care and other clinical settings. Our results emphasise the importance of using these measures alongside clinical assessments that take in the perspective of the patient. Sole reliance upon information from self-administered questionnaires can potentially be misleading and ignores areas that patients' regard as important. Our evidence supports the widespread scepticism among physicians about using self-administered questionnaires in clinical practice [64]. We provide quantitative evidence that the results of these

questionnaires need to be interpreted along with other clinical assessments and should not be relied upon alone. Our findings support the concept of 'personal recovery', developed in mental health services but also relevant in primary care [77,78]. Personal recovery emphasizes the importance of a holistic focus on patients' broad experiences rather than a restricted focus on 'clinical recovery' or symptom change. This makes the patients' voice of central importance and there are efforts under-way to devise better measurements of patient-reported recovery.

Some patients view self-administered questionnaires positively and request them to monitor their recovery [79]. Questionnaires can, therefore, play a useful role in outcome assessment, in conjunction with clinical assessment that takes account of more holistic changes in mood. They are also useful as a guide for service level outcome assessment [61]. In clinical trials, self-administered questionnaires are widely used for comparing groups and such randomized comparisons should be unbiased. Our findings suggest, though, that additional questions should also be used to assess the outcome of treatments in research studies.

Future research could examine the generalizability of our findings to international settings and mental health services, and the relationship between patient-rated change and other mental health measures including the outcomes used in the NHS Improving Access to Psychological Therapy (IAPT) services [61]. Future clinical trials could also use the patient-rated change in mood question as an outcome that might help to address the limitations of existing measures.

Table 1: Sample characteristics at baseline.

Demographic Variable	Overall Sample (n = 554)
Age, mean (SD)	48·30 (12·56)
Female, N (%)	377 (68)
White, N (%)	530 (96)
Married or partnership, N (%)	278 (50)
Employed, N (%)	296 (53)
Higher Education, N (%)	161 (29)
ICD-10 Depression Diagnosis, N (%)	238 (45)
Taking antidepressants, N (%)	377 (69)
Site	
Bristol	197 (36)
Liverpool	188 (34)
York	169 (30)

Table 2: Change in depression severity according to the patient-rated change scale, compared to clinically important changes in PHQ-9 and BDI-II scores. Disagreement (differing indications of change in depressive symptoms) is shaded in grey (n = 465 PHQ-9, n = 468 BDI-II).

	Patient-rated change scale				
	Feeling a lot better	Feeling slightly better	Feeling about the same	Feeling slightly worse	Feeling a lot worse
PHQ-9					
Mean (SD) change	–3·4 (4·1)	–2·7 (3·9)	–·26 (3·6)	1·3 (4·3)	1·6 (5·4)
CID Decrease, n (%)[a]	19 (47%)	34 (35%)	29 (14%)	9 (9%)	2 (13%)
No CID Change, n (%)[a]	20 (50%)	56 (58%)	149 (70%)	65 (66%)	11 (69%)
CID Increase, n (%)[a]	1 (3%)	7 (7%)	36 (16%)	24 (25%)	3 (18%)
Disagreement, n (%)[b]	21 (53%)	63 (65%)	65 (30%)	74 (75%)	13 (82%)
BDI-II					
Mean (SD) change	–8·0 (8·9)	–5·6 (6·5)	–1·2 (5·8)	0·0 (5·7)	3·2 (7·1)
CID Decrease, n (%)[a]	29 (72%)	55 (57%)	74 (34%)	21 (22%)	3 (18%)
No CID Change, n (%)[a]	9 (23%)	33 (34%)	92 (42%)	48 (49%)	9 (53%)
CID Increase, n (%)[a]	2 (5%)	9 (9%)	51 (24%)	28 (29%)	5 (29%)
Disagreement, n (%)[b]	11 (28%)	42 (43%)	125 (58%)	69 (71%)	12 (71%)

CID = Clinically Important Difference based on the Minimal CID (MCID).

[a]Percentages represent the proportions of patients showing differing CID changes (decrease, no change, increase) within each category of the global rating of change scale.

[b]Percentages represent the proportions of patients showing disagreement within each category of the global rating of change scale

Table 3. Association between exposure variables and the odds of reporting feeling better (versus the same or worse), adjusted for change on the PHQ-9

Exposure variable	Odds ratio for reporting feeling better (versus the same or worse), 95% confidence interval and p value (n=375)		
Anxiety symptoms[a]	**Unadjusted**		
	Baseline to follow-up 1	Follow-up 1 to 2	Follow-up 3 to 4
Feeling same or worse	ref	ref	ref
Feeling better	·67 (·55 to ·82) ·<.0001	·65 (·53 to ·79) <·0001	·71 (·59 to .86) <·0001
	Adjusted[d]		
Feeling same or worse	ref	ref	ref
Feeling better	·66 (.54 to ·82) ·016	·61 (·49 to ·76) <·0001	.72 (.60 to .97) ·001
Mental health related quality of life[a]	**Undjusted**		
Feeling same or worse	ref	ref	ref
Feeling better	1·34 (1·11 to 1·61) ·002	1·33 (1·11 to 1·59) ·002	1·38 (1·15 to 1·64) ·000
	Adjusted[d]		
Feeling same or worse	ref	ref	Ref
Feeling better	1·32 (1·08 to 1·61) ·006	1·38 (1·14 to 1·66) ·001	1·40 (1·17 to 1·68) <·000
Physical health related quality of life[a]	**Unadjusted**		
Feeling same or worse	ref	ref	Ref
Feeling better	1·28 (1·07 to 1·54) ·007	1·25 (1·06 to 1·48) ·009	1·20 (1·01 to 1·42) ·039
	Adjusted[d]		
Feeling same or worse	Ref	Ref	Ref
Feeling better	1·32 (1·08 to 1·60) ·006	1·32 (1·10 to 1·58) ·002	1·19 (·99 to 1·43) ·057
Negative life events[b]	**Unadjusted**		
Feeling same or worse	ref	ref	ref
Feeling better	·98 (·61 to 1·59) ·94	1·13 (·72 to 1·79) ·59	1·17 (·74 to 1·85) ·50
	Adjusted[d]		
Feeling same or worse	Ref	Ref	Ref
Feeling better	·99 (·60 to 1·65) ·98	1·11 (·69 to 1·78) ·76	1·15 (·72 to 1·85) ·56
Social support[c]	**Unadjusted odds Ratio (95% CI) p value**		
Feeling same or worse	Ref	Ref	ref
Feeling better	1·07 (1·00 to 1·14) ·067	1·01 (·95 to 1·07) ·71	1·02 (·96 to 1·08) ·56
	Adjusted[d]		
Feeling same or worse	Ref	Ref	Ref
Feeling better	1·07 (1·00 to 1·15) ·045	1·02 (·96 to 1·08) ·59	1·01 (·95 to 1·08) ·76

[a]For exposures measured at every time-point (anxiety and quality of life), odds ratios represent the odds of reporting feeling better for each four-point increase in anxiety symptoms over time (on a factor score obtained using principal components analysis), adjusted for a binary indicator of meaningful change on the PHQ9.

[b]Negative life events was measured at baseline only. The odds ratio represents the odds of feeling better in those who reported one life event or more compared to those who reported no life events, adjusted for a binary indicator of meaningful change on the PHQ9.

[c]Social support was measured at baseline only. The odds ratio represents the odds of reporting feeling better for each standard deviation increase in social support, adjusted for a binary indicator of meaningful change on the PHQ9.

[d]Adjusted for age, sex, ethnicity, site, education level, current use of antidepressants and marital, financial and employment status

Table 4. Association between exposure variables and the odds of reporting feeling better (versus the same or worse), adjusted for change on the BDI-II

Exposure variable	Odds ratio for reporting feeling better (versus the same or worse), 95% confidence interval and p value (n=375)		
Anxiety symptoms[a]	**Unadjusted**		
	Baseline to follow-up 1	Follow-up 1 to 2	Follow-up 3 to 4
Feeling same or worse	ref	ref	ref
Feeling better	·67 (·56 to ·81) <·0001	·67 (·56 to ·81) <·0001	·70 (·59 to ·84) <·0001
	Adjusted[c]		
Feeling same or worse	Ref	ref	Ref
Feeling better	·65 (·53 to ·81) <·0001	·61 (·49 to ·76) <·0001	·71 (·59 to ·86) <·0001
Mental health related quality of life[a]	**Undjusted**		
Feeling same or worse	ref	ref	ref
Feeling better	1·37 (1·13 to 1·65) ·001	1·33 (1·12 to 1·58) ·001	1·38 (1·16 to 1·64) <·0001
	Adjusted[d]		
Feeling same or worse	ref	ref	ref
Feeling better	1·34 (1·10 to 1·63) ·004	1·38 (1·14 to 1·66) ·001	1·38 (1·16 to 1·64) <·0001
Physical health related quality of life[a]	**Unadjusted**		
Feeling same or worse	ref	ref	ref
Feeling better	1·25 (1·04 to 1·49) ·016	1·24 (1·05 to 1·46) ·013	1·22 (1·03 to 1·45) ·021
	Adjusted[d]		
Feeling same or worse	Ref	Ref	Ref
Feeling better	1·27 (1·05 to 1·54) ·015	1·30 (1·08 to 1·55) ·005	1·22 (1·02 to 1·47) ·030
Negative life events[b]	**Unadjusted**		
Feeling same or worse	ref	ref	Ref
Feeling better	1·03 (·64 to 1·66) ·89	1·18 (·75 to 1·85) ·49	1·14 (·71 to 1·81) ·59
	Adjusted[d]		
Feeling same or worse	ref	Ref	Ref
Feeling better	1·04 (·63 to 1·72) ·87	1·15 (·72 to 1·85) ·55	1·11 (·68 to 1·79) ·68
Social support[c]	**Unadjusted**		
Feeling same or worse	ref	Ref	Ref
Feeling better	1·07 (1· to 1·14) ·06	1·01 (·95 to 1·07) ·71	1·02 (·96 to 1·09) ·52
	Adjusted[d]		
Feeling same or worse	ref	Ref	Ref
Feeling better	1·07 (1·00 to 1·15) ·044	1·02 (·96 to 1·08) ·59	1·01 (·95 to 1·08) ·70

[a]For exposures measured at every time-point (anxiety and quality of life), odds ratios represent the odds of reporting feeling better for each four-point increase in anxiety symptoms over time (on a factor score obtained using principal components analysis), adjusted for a binary indicator of meaningful change on the PHQ9.

[b]Negative life events was measured at baseline only. The odds ratio represents the odds of feeling better in those who reported one life event or more compared to those who reported no life events, adjusted for a binary indicator of meaningful change on the PHQ9.

[c]Social support was measured at baseline only. The odds ratio represents the odds of reporting feeling better for each standard deviation increase in social support, adjusted for a binary indicator of meaningful change on the PHQ9.

[d]Adjusted for age, sex, ethnicity, site, education level, current use of antidepressants and marital, financial and employment status

References

1. WHO. *fact sheets*. (2018).

2. NHS digital. *Prescription Cost Analysis*. (2018).

3. Hamilton, M. A rating scale for depression. *J. Neurol. Neurosurg. Psychiatry* **23,** 56–62 (1960).

4. *Depression: management of depression in primary and secondary care. Clinical Guideline 23*. (National Institute for Clinical Excellence, 2004).

5. Jacobson, N. S. & Truax, P. Clinical Significance: A Statistical Approach to Defining Meaningful Change in Psychotherapy Research. *J. Consult. Clin. Psychol.* **59,** 12–19 (1991).

6. McMillan, D., Gilbody, S. & Richards, D. Defining successful treatment outcome in depression using the PHQ-9: A comparison of methods. *J Affect Disord* **127,** 122–129 (2010).

7. Jaeschke, R., Singer, J. & Guyatt, G. H. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* **10,** 407–415 (1989).

8. Gilbody, S., Richards, D., Brealey, S. & Hewitt, C. Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. *J Gen Intern Med* **22,** 1596–1602 (2007).

9. Excellence, N. I. for H. and C. *Depression (updated edition)*. (The British Psychological Society and The Royal College of Psychiatrists, 2009).

10. Lowe, B., Unutzer, J., Callahan, C. M., Perkins, A. J. & Kroenke, K. Monitoring depression treatment outcomes with the patient health questionnaire-9. *Med.Care* **42,** 1194–1201 (2004).

11. Jacobson, N. & Greenley, D. What Is Recovery? A Conceptual Model and Explication. *Psychiatr. Serv.* **52,** 482–485 (2001).

12. Ridge, D. & Ziebland, S. 'The old me could never have done that': how people give meaning to recovery following depression. *Qual Heal. Res* **16,** 1038–1053 (2006).

13. Malpass, A., Shaw, A., Kessler, D. & Sharp, D. Concordance between PHQ-9 scores and patients' experiences of depression: A mixed methods study. *Br. J. Gen. Pract.* **60,** 231–8 (2010).

14. Harmer, C. J., Goodwin, G. M. & Cowen, P. J. Why do antidepressants take so long to work? A cognitive neuropsychological model of antidepressant drug action. *Br J Psychiatry* **195,** 102–108 (2009).

15. Beck, A T Steer, R A, Brown, G. K. *Beck Depression Inventory Manual*. (Psychological Corporation, 1996).

16. Zigmond, A. & Snaith, R. The hospital anxiety and depression scale. *Acta Psychiatr. Scand.* **67,** 361–370 (1983).

17. Cameron, I. M., Crawford, J. R., Lawton, K. & Reid, I. C. Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care. *Br J Gen Pr.* **58,** 32–36 (2008).

18. NICE. NICE. Management of depression in primary and secondary care: Clinical Guidelines 23. (2010).

19. Kirsch, I. *et al.* Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. *PLoS Med* **5,** e45 (2008).

20. Khan, A., Leventhal, R. M., Khan, S. R. & Brown, W. A. Severity of depression and response to antidepressants and placebo: an analysis of the Food and Drug Administration database. *J.Clin.Psychopharmacol.* **22,** 40–45 (2002).

21. Fournier, J. C. *et al.* Antidepressant Drug effects and Depression Severity: A Patient- Level Meta-Analysis. *JAMA* **6,** 47–53 (2010).

22. Gibbons, R. D., Hur, K., Hendricks Brown, C., Davis, J. M. & Mann, J. J. Who Benefits from Antidepressants?: Synthesis of 6-Week Patient-Level Outcomes from Double-Blind Placebo Controlled Randomized Trials of Fluoxetine and Venlafaxine. *Arch Gen Psychiatry* **69,** 572–579 (2012).

23. Rabinowitz, J. *et al.* Initial depression severity and response to antidepressants v. placebo: patient-level data analysis from 34 randomised controlled trials. *Br. J. Psychiatry* **209,** 427–428 (2016).

24. Furukawa, T. A. *et al.* Initial severity of major depression and efficacy of new generation antidepressants: individual participant data meta-analysis. *Acta Psychiatr. Scand.* **137,** 450–458 (2018).

25. Barbui, C., Cipriani, A., Patel, V., yuso-Mateos, J. L. & van, O. M. Efficacy of antidepressants and benzodiazepines in minor depression: systematic review and meta-analysis. *Br J Psychiatry* **198,** 11–6, sup (2011).

26. Zimmerman, M., Posternak, M. A. & Chelminski, I. Symptom severity and exclusion from antidepressant efficacy trials. *J. Clin. Psychopharmacol.* **22,** 610–4 (2002).

27. Cipriani, A. *et al.* Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *Lancet (London, England)* **0,** (2018).

28. World Health Organization. *Classification of Mental and Behavioural Disorders*. *Geneva: World Health Organisation* (1992).

29. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition. *Washingt. DC Am. Psychiatr. Assoc.* **10,** 943 (1994).

30. Rai, D., Skapinakis, P., Wiles, N., Lewis, G. & Araya, R. Common mental disorders, subthreshold symptoms and disability: Longitudinal study. *Br. J. Psychiatry* **197,** (2010).

31. Broadhead, W. E., Blazer, D. G., George, L. K. & Tse, C. K. Depression, Disability Days, and Days Lost From Work in a Prospective Epidemiologic Survey. *JAMA J. Am. Med. Assoc.* **264,** 2524–2528 (1990).

32. de Lima, M. S., Hotoph, M. & Wessely, S. The efficacy of drug treatments for dysthymia: a systematic review and meta-analysis. *Psychol. Med.* **29,** 1273–89 (1999).

33. de Lima, M. S. Review: antidepressant drugs are effective in dysthymia.

34. Anderson, I. M., Nutt, D. J. & Deakin, J. F. Evidence-based guidelines for treating depressivedisorders with antidepressants: a revision of the 1993 British Association for Psychopharmacology guidelines. British Association for Psychopharmacology. *J Psychopharmacol* **14(1),** 3–20 (2000).

35. Salaminios, G. *et al.* A randomised controlled trial assessing the severity and duration of depressive symptoms associated with a clinically significant response to sertraline versus placebo, in people presenting to primary care with depression (PANDA trial): study protocol for. *Trials* **18,** 496 (2017).

36. Thomas, L. *et al.* GENetic and clinical Predictors Of treatment response in Depression: the GenPod randomised trial protocol. *Trials* **9,** 29 (2008).

37. Thomas, L. J. *et al.* Cognitive behavioural therapy as an adjunct to pharmacotherapy for treatment resistant depression in primary care: The CoBalT randomised controlled trial protocol. *Contemp. Clin. Trials* (2012). doi:10.1016/j.cct.2011.10.016

38. Baxter, H. *et al.* Physical activity as a treatment for depression: the TREAD randomised trial protocol. *Trials* **11,** 105 (2010).

39. Montgomery, S. A. & Asberg, M. A new depression scale designed to be sensitive to change. *Br. J. Psychiatry* **134,** 382–9 (1979).

40. Brooks, R. & EuroQol, G. Euroqol: the current state of play. *Health Policy (New. York).* **37,** 53–72 (1996).

41. Stewart, A. D., Hays, R. D. & Ware, J. E. The MOS short-form General Health Survey. *Med. Care* **26,** 724–732 (1988).

42. Lu, G., Kounali, D. & Ades, A. E. Simultaneous Multioutcome Synthesis and Mapping of Treatment Effects to a Common Scale. *Value Heal.* **17,** 280–287 (2014).

43. Titov, N. *et al.* Psychometric comparison of the PHQ-9 and BDI-II for measuring response during treatment of depression. *Cogn. Behav. Ther.* (2011). doi:10.1080/16506073.2010.550059

44. Fournier, J. C. *et al.* Antidepressant drug effects and depression severity: A patient-level meta-analysis. *JAMA - Journal of the American Medical Association* **303,** 47–53 (2010).

45. Gibbons, R. D., Hur, K., Brown, C. H., Davis, J. M. & Mann, J. J. Benefits from antidepressants: synthesis of 6-week patient-level outcomes from double-blind placebo-controlled randomized trials of fluoxetine and venlafaxine. *Arch Gen Psychiatry* **69,** 572–579 (2012).

46. Cipriani, A. *et al.* Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet* **373,** 746–758 (2009).

47. Lunn, D., Jackson, C., Best, N., Thomas, A. & Spiegelhalter, D. *The BUGS book: A practical introduction to Bayesian analysis*. (2013).

48. Briggs AH, Claxton K, S. M. *Decision modelling for health economic evaluation.* (Oxford University Press, 2006).

49. Welton NJ, Sutton AJ, Cooper NJ, Abrams KR, A. A. *Evidence Synthesis for Decision Making in Healthcare*. (John Wiley and Sons, 2012).

50. Ades, A. E., Lu, G. & Claxton, K. Expected Value of Sample Information Calculations in Medical Decision Modeling. *Med. Decis. Mak.* **24,** 207–227 (2004).

51. Lewis, G., Pelosi, A. J., Araya, R. & Dunn, G. Measuring psychiatric disorder in the community: A standardized assessment for use by lay interviewers. *Psychol. Med.* **22,** (1992).

52. Ritchie, J. & Spencer, L. in *The Qualitative Research Companion* 305–329 (2002).

53. Willis, G. B. *Cognitive interviewing : a tool for improving questionnaire design*. *Techniques* (2005). doi:http://dx.doi.org/10.4135/9781412983655

54. Faria, R., Gomes, M., Epstein, D. & White, I. R. A Guide to Handling Missing Data in Cost-Effectiveness Analysis Conducted Within Randomised Controlled Trials. *Pharmacoeconomics* **32,** 1157–1170 (2014).

55. Sterne, J. A. C. *et al.* Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ (Online)* **339,** 157–160 (2009).

56. Hoch, J. S., Briggs, A. H. & Willan, A. R. Something old, something new, something borrowed, something blue: A framework for the marriage of health econometrics and cost-effectiveness analysis. *Health Econ.* **11,** 415–430 (2002).

57. Hoch, J. S. & Dewa, C. S. Advantages of the net benefit regression framework for economic evaluations of interventions in the workplace: A case study of the cost-effectiveness of a collaborative mental health care program for people receiving short-term disability benefits for psychiatric disorders. *J. Occup. Environ. Med.* (2014). doi:10.1097/JOM.0000000000000130

58. Kounali, D. Z., Button, K. S., Lewis, G. & Ades, A. E. The relative responsiveness of test instruments can be estimated using a meta-analytic approach: an illustration with treatments for depression. *J Clin Epidemiol* (2016). doi:10.1016/j.jclinepi.2016.03.005

59. Kendrick, T. *et al.* Management of depression in UK general practice in relation to scores on depression severity questionnaires: analysis of medical record data. *BMJ* **338,** b750 (2009).

60. Thombs, B. D. & Ziegelstein, R. C. Does depression screening improve depression outcomes in primary care? *BMJ* **348,** g1253 (2014).

61. Clark, D. M. *et al.* Transparency about the outcomes of mental health services (IAPT approach): an analysis of public data. *Lancet (London, England)* **391,** 679–686 (2018).

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

63. Beck, A. T., Guth, D., Steer, R. A. & Ball, R. Screening for major depression disorders in medical inpatients with the Beck Depression Inventory for Primary Care. *Behav. Res. Ther.* **35,** 785–91 (1997).

64. Dowrick, C. *et al.* Patients' and doctors' views on depression severity questionnaires incentivised in UK quality and outcomes framework: qualitative study. *BMJ* **338,** b663 (2009).

65. Toop, L. The QOF, NICE, and depression: a clumsy mechanism that undermines clinical judgment. *Br. J. Gen. Pract.* **61,** 432–3 (2011).

66. Robinson, J. *et al.* Why are there discrepancies between depressed patients' Global Rating of Change and scores on the Patient Health Questionnaire depression module? A qualitative study of primary care in England. *BMJ Open* **7,** (2017).

67. Malpass, A. *et al.* Usefulness of PHQ-9 in primary care to determine meaningful symptoms of low mood: a qualitative study. *Br. J. Gen. Pract.* **66,** e78-84 (2016).

68. Kamper, S. J., Maher, C. G. & Mackay, G. Global rating of change scales: a review of strengths and weaknesses and considerations for design. *J. Man. Manip. Ther.* **17,** 163–70 (2009).

69. Fischer, D. *et al.* Capturing the Patient's View of Change as a Clinical Outcome Measure. *JAMA* **282,** 1157 (1999).

70. Kroenke, K., Spitzer, R. L. & Williams, J. B. The PHQ-9: validity of a brief depression severity measure. *J. Gen. Intern. Med.* **16,** 606–13 (2001).

71. Button, K. S. *et al.* Minimal clinically important difference on the Beck Depression Inventory-II according to the patient's perspective. *Psychol. Med.* **45,** 3269–3279 (2015).

72. Spitzer, R. L., Kroenke, K., Williams, J. B. W. & Löwe, B. A Brief Measure for Assessing Generalized Anxiety Disorder. *Arch. Intern. Med.* **166,** 1092 (2006).

73. Ware, J., Kosinski, M. & Keller, S. D. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. *Med. Care* **34,** 220–33 (1996).

74. Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Philos. Trans. A. Math. Phys. Eng. Sci.* **374,** 20150202 (2016).

75. Herrmann, D. Reporting current, past, and changed health status. What we know about distortion. *Med. Care* **33,** AS89-94 (1995).

76. Schwartz, C. E. & Sprangers, M. A. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Soc. Sci. Med.* **48,** 1531–48 (1999).

77. Bejerholm, U. & Roe, D. Personal recovery within positive psychiatry. *Nord. J. Psychiatry* 1–11 (2018). doi:10.1080/08039488.2018.1492015

78. Burgess, P., Pirkis, J., Coombs, T. & Rosen, A. Assessing the Value of Existing Recovery Measures for Routine Use in Australian mental Health Services. *Aust. New Zeal. J. Psychiatry* **45,** 267–280 (2011).

79. Moore, M. *et al.* Depression management in primary care: an observational study of management changes related to PHQ-9 score for depression monitoring. *Br. J. Gen. Pract.* **62,** e451–e457 (2012).

80. Mathers, C. D. & Loncar, D. Projections of Global Mortality and Burden of Disease from 2002 to 2030. *PLoS Med.* **3,** e442 (2006).

81. NHS Digital. *Prescriptions Dispensed in the Community - Statistics for England, 2006-2016*.

82. Linde, K. *et al.* Efficacy and acceptability of pharmacological treatments for depressive disorders in primary care: systematic review and network meta-analysis. *Ann. Fam. Med.* **13,** 69–79 (2015).

83. Cameron, I. M. *et al.* Measuring depression severity in general practice: discriminatory performance of the PHQ-9, HADS-D, and BDI-II. *Br. J. Gen. Pract.* **61,** e419-26 (2011).

84. Cuijpers, P., de Graaf, R. & van Dorsselaer, S. Minor depression: risk profiles, functional disability, health care use and risk of developing major depression. *J. Affect. Disord.* **79,** 71–79 (2004).

85. Simon, G. E. *et al.* Antidepressants are not overprescribed for mild depression. *J. Clin. Psychiatry* **76,** 1627–32 (2015).

86. Kirsch, I. *et al.* Initial Severity and Antidepressant Benefits: A Meta-Analysis of Data Submitted to the Food and Drug Administration. *PLoS Med.* **5,** e45 (2008).

87. Khan, A., Leventhal, R. M., Khan, S. R. & Brown, W. A. Severity of depression and response to antidepressants and placebo: an analysis of the Food and Drug Administration database. *J. Clin. Psychopharmacol.* **22,** 40–5 (2002).

88. Rabinowitz, J. *et al.* Initial depression severity and response to antidepressants v. placebo: patient-level data analysis from 34 randomised controlled trials. *Br. J. Psychiatry* **209,** 427– 428 (2016).

89. Barbui, C., Cipriani, A., Patel, V., Ayuso-Mateos, J. L. & van Ommeren, M. Efficacy of antidepressants and benzodiazepines in minor depression: systematic review and meta-analysis. *Br. J. Psychiatry* **198,** 11–16 (2011).

90. Cameron, I. M., Reid, I. C. & MacGillivray, S. A. Efficacy and tolerability of antidepressants for sub-threshold depression and for mild major depressive disorder. *J. Affect. Disord.* **166,** 48–58 (2014).

91. Baldwin, D. *et al.* Evidence-based guidelines for treating depressive disorders with antidepressants: A revision of the 2008 British Association for Psychopharmacology guidelines. *J. Psychopharmacol.* **29,** 459–525 (2015).

92. Peto, R. & Baigent, C. Trials: the next 50 years. Large scale randomised evidence of moderate benefits. *BMJ* **317,** 1170–1 (1998).

93. Freedman, B. Equipoise and the Ethics of Clinical Research. *N. Engl. J. Med.* **317,** 141–145 (1987).

94. Cipriani, A. *et al.* Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet (London, England)* **373,** 746–58 (2009).

95. Lewis, G., Pelosi, A. J., Araya, R. & Dunn, G. Measuring psychiatric disorder in the community: a standardized assessment for use by lay interviewers. *Psychol. Med.* **22,** 465 (1992).

96. Hotopf, M., Lewis, G. & Normand, C. Putting trials on trial--the costs and consequences of small trials in depression: a systematic review of methodology. *J. Epidemiol. Community Health* **51,** 354–8 (1997).

97. Lewis, G. Observer bias in the assessment of anxiety and depression. *Soc. Psychiatry Psychiatr. Epidemiol.* **26,** 265–272 (1991).

98. Titov, N. *et al.* Psychometric Comparison of the PHQ-9 and BDI-II for Measuring Response during Treatment of Depression. *Cogn. Behav. Ther.* **40,** 126–136 (2011).

99. Crawford, A. A. *et al.* Adverse effects from antidepressant treatment: randomised controlled trial of 601 depressed individuals. *Psychopharmacology (Berl).* **231,** 2921–2931 (2014).

100. Angst, F., Aeschlimann, A. & Angst, J. The minimal clinically important difference raised the significance of outcome effects above the statistical level, with methodological implications for future studies. *J. Clin. Epidemiol.* **82,** 128–136 (2017).

101. Sterne, J. A. C. *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* **338,** b2393 (2009).

102. McManus, S., Bebbington, P., Jenkins, R. & Brugha, T. *Mental health and wellbeing in England: Adult Psychiatric Morbidity Survey 2014*. (NHS Digital, 2016).

103. Baldwin, D., Woods, R., Lawson, R. & Taylor, D. Efficacy of drug treatments for generalised anxiety disorder: systematic review and meta-analysis. *BMJ* **342,** d1199 (2011).

104. Harmer, C. J., Duman, R. S. & Cowen, P. J. How do antidepressants work? New perspectives for refining future treatment approaches. *The lancet. Psychiatry* **4,** 409–418 (2017).

105. Slee, A. *et al.* Pharmacological treatments for generalised anxiety disorder: a systematic review and network meta-analysis. *Lancet (London, England)* **0,** (2019).

106. Wittchen, H.-U. *et al.* Generalized anxiety and depression in primary care: prevalence, recognition, and management. *J. Clin. Psychiatry* **63 Suppl 8,** 24–34 (2002).