NIHR | National Institute for Health Research

Journals Library

Check for updates

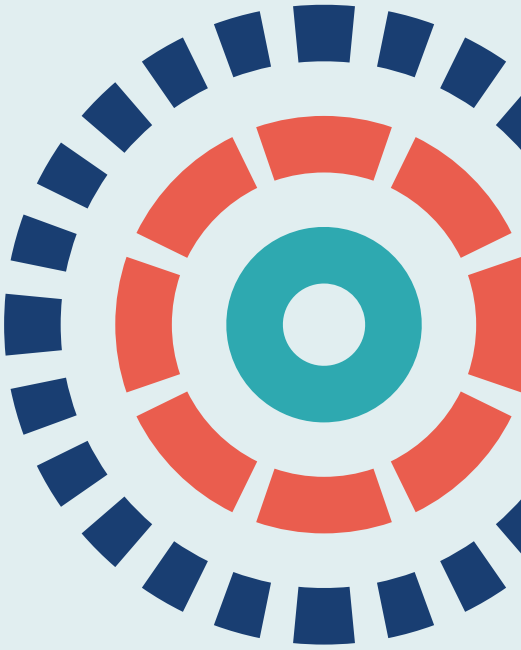
Programme Grants for Applied Research

*Volume 7 • Issue 10 • December 2019*

*ISSN 2050-4322*

# Antidepressant treatment with sertraline for adults with depressive symptoms in primary care: the PANDA research programme including RCT

*Larisa Duffy, Gemma Lewis, Anthony Ades, Ricardo Araya, Jessica Bone, Sally Brabyn, Katherine Button, Rachel Churchill, Tim Croudace, Catherine Derrick, Padraig Dixon, Christopher Dowrick, Christopher Fawsitt, Louise Fusco, Simon Gilbody, Catherine Harmer, Catherine Hobbs, William Hollingworth, Vivien Jones, Tony Kendrick, David Kessler, Naila Khan, Daphne Kounali, Paul Lanham, Alice Malpass, Marcus Munafo, Jodi Pervin, Tim Peters, Derek Riozzie, Jude Robinson, George Salaminios, Debbie Sharp, Howard Thom, Laura Thomas, Nicky Welton, Nicola Wiles, Rebecca Woodhouse and Glyn Lewis*



*DOI 10.3310/pgfar07100*

# Antidepressant treatment with sertraline for adults with depressive symptoms in primary care: the PANDA research programme including RCT

Larisa Duffy,[1]* Gemma Lewis,[1] Anthony Ades,[2] Ricardo Araya,[3] Jessica Bone,[1] Sally Brabyn,[4] Katherine Button,[5] Rachel Churchill,[6] Tim Croudace,[7] Catherine Derrick,[2] Padraig Dixon,[2] Christopher Dowrick,[8] Christopher Fawsitt,[2] Louise Fusco,[8] Simon Gilbody,[4] Catherine Harmer,[9] Catherine Hobbs,[5] William Hollingworth,[2] Vivien Jones,[2] Tony Kendrick,[10] David Kessler,[2] Naila Khan,[8] Daphne Kounali,[2] Paul Lanham,[11] Alice Malpass,[2] Marcus Munafo,[12] Jodi Pervin,[4] Tim Peters,[2] Derek Riozzie,[11] Jude Robinson,[13] George Salaminios,[1] Debbie Sharp,[2] Howard Thom,[2] Laura Thomas,[2] Nicky Welton,[2] Nicola Wiles,[2] Rebecca Woodhouse[4] and Glyn Lewis[1]

[1]Division of Psychiatry, University College London, London, UK
[2]Bristol Medical School, University of Bristol, Bristol, UK
[3]Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
[4]Department of Health Sciences, University of York, York, UK
[5]Department of Psychology, University of Bath, Bath, UK
[6]Centre for Reviews and Dissemination, University of York, York, UK
[7]School of Nursing and Health Studies, University of Dundee, Dundee, UK
[8]Institute of Psychology Health and Society, University of Liverpool, Liverpool, UK
[9]Department of Psychiatry, University of Oxford, Oxford, UK
[10]Primary Care and Population Sciences, Faculty of Medicine, University of Southampton, Southampton, UK
[11]Patient and public involvement contributor, UK
[12]Department of Psychology and Integrated Epidemiology Unit, University of Bristol, Bristol, UK
[13]Department of Sociology, Social Policy and Criminology, University of Liverpool, Liverpool, UK

*Corresponding author

**Declared competing interests of authors:** Rachel Churchill reports grants from the National Institute for Health Research (NIHR) (Programme Grants for Applied Research programme) during the conduct of the study. Simon Gilbody serves as deputy chairperson of the NIHR Health Technology Commissioning Board, but was not involved in the commissioning of this programme of research. Catherine Harmer reports personal fees from P1vital (Wallingford, UK), grants from UCB Pharma (Brussels, Belgium), grants and personal fees from Johnson & Johnson (New Brunswick, NJ, USA), and personal fees from H. Lundbeck A/S (Copenhagen, Denmark), Servier Laboratories (Neuilly-sur-Seine, France) and Pfizer Inc. (New York, NY, USA) outside the submitted work. Tony Kendrick reports grants from NIHR during the conduct of the study. Marcus Munafo reports grants and personal fees from Cambridge Cognition (Cambridge, UK) and personal fees from Jericoe Ltd (Bristol, UK) outside the submitted work. Tim Peters reports grants from NIHR during the conduct of the study. Howard Thom reports personal fees from Novartis Pharma AG (Basel, Switzerland), Pfizer Inc., Roche Holding AG (Basel, Switzerland) and Eli Lilly and Company (Indianapolis, IN, USA) outside the submitted work. Nicky Welton reports grants from NIHR during the conduct of the study; and she is the principal investigator on a Medical Research Council-funded project in collaboration with Pfizer Inc. Pfizer Inc. part funded a junior researcher on the project. The project is purely methodological using historical data on pain relief, and unrelated to this work. Nicola Wiles reports grants from NIHR during the conduct of the study. Glyn Lewis reports grants from University College London during the conduct of the study and personal fees from Fortitude Law (London, UK) outside the submitted work.

Published December 2019
DOI: 10.3310/pgfar07100

This report should be referenced as follows:

Duffy L, Lewis G, Ades A, Araya R, Bone J, Brabyn S, *et al*. Antidepressant treatment with sertraline for adults with depressive symptoms in primary care: the PANDA research programme including RCT. *Programme Grants Appl Res* 2019;**7**(10).

# Programme Grants for Applied Research

ISSN 2050-4322 (Print)

ISSN 2050-4330 (Online)

This journal is a member of and subscribes to the principles of the Committee on Publication Ethics (COPE) (www.publicationethics.org/).

Editorial contact: journals.library@nihr.ac.uk

The full PGfAR archive is freely available to view online at www.journalslibrary.nihr.ac.uk/pgfar. Print-on-demand copies can be purchased from the report pages of the NIHR Journals Library website: www.journalslibrary.nihr.ac.uk

**Criteria for inclusion in the *Programme Grants for Applied Research* journal**
Reports are published in *Programme Grants for Applied Research* (PGfAR) if (1) they have resulted from work for the PGfAR programme, and (2) they are of a sufficiently high scientific quality as assessed by the reviewers and editors.

## Programme Grants for Applied Research programme

The Programme Grants for Applied Research (PGfAR) programme, part of the National Institute for Health Research (NIHR), was established in 2006 to fund collaborative, multidisciplinary programmes of applied research to solve health and social care challenges. Findings are expected to provide evidence that lead to clear and identifiable patient benefits, in the relatively near future.

PGfAR is researcher led and does not specify topics for research; however, the research must be in an area of priority or need for the NHS and the social care sector of the Department of Health and Social Care, with particular emphasis on health and social care areas that cause significant burden, where other research funders may not be focused, or where insufficient funding is available.

The programme is managed by the NIHR Central Commissioning Facility (CCF) with strategic input from the Programme Director.
For more information about the PGfAR programme please visit the website: https://www.nihr.ac.uk/explore-nihr/funding-programmes/programme-grants-for-applied-research.htm

## This report

The research reported in this issue of the journal was funded by PGfAR as project number RP-PG-0610-10048. The contractual start date was in March 2012. The final report began editorial review in September 2018 and was accepted for publication in June 2019. As the funder, the PGfAR programme agreed the research questions and study designs in advance with the investigators. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The PGfAR editors and production house have tried to ensure the accuracy of the authors' report and would like to thank the reviewers for their constructive comments on the final report document. However, they do not accept liability for damages or losses arising from material published in this report.

This report presents independent research funded by the National Institute for Health Research (NIHR). The views and opinions expressed by authors in this publication are those of the authors and do not necessarily reflect those of the NHS, the NIHR, CCF, NETSCC, PGfAR or the Department of Health and Social Care. If there are verbatim quotations included in this publication the views and opinions expressed by the interviewees are those of the interviewees and do not necessarily reflect those of the authors, those of the NHS, the NIHR, NETSCC, the PGfAR programme or the Department of Health and Social Care.



Published by the NIHR Journals Library (www.journalslibrary.nihr.ac.uk), produced by Prepress Projects Ltd, Perth, Scotland (www.prepress-projects.co.uk).

# Abstract

## Antidepressant treatment with sertraline for adults with depressive symptoms in primary care: the PANDA research programme including RCT

Larisa Duffy,[1]* Gemma Lewis,[1] Anthony Ades,[2] Ricardo Araya,[3] Jessica Bone,[1] Sally Brabyn,[4] Katherine Button,[5] Rachel Churchill,[6] Tim Croudace,[7] Catherine Derrick,[2] Padraig Dixon,[2] Christopher Dowrick,[8] Christopher Fawsitt,[2] Louise Fusco,[8] Simon Gilbody,[4] Catherine Harmer,[9] Catherine Hobbs,[5] William Hollingworth,[2] Vivien Jones,[2] Tony Kendrick,[10] David Kessler,[2] Naila Khan,[8] Daphne Kounali,[2] Paul Lanham,[11] Alice Malpass,[2] Marcus Munafo,[12] Jodi Pervin,[4] Tim Peters,[2] Derek Riozzie,[11] Jude Robinson,[13] George Salaminios,[1] Debbie Sharp,[2] Howard Thom,[2] Laura Thomas,[2] Nicky Welton,[2] Nicola Wiles,[2] Rebecca Woodhouse[4] and Glyn Lewis[1]

[1]Division of Psychiatry, University College London, London, UK
[2]Bristol Medical School, University of Bristol, Bristol, UK
[3]Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
[4]Department of Health Sciences, University of York, York, UK
[5]Department of Psychology, University of Bath, Bath, UK
[6]Centre for Reviews and Dissemination, University of York, York, UK
[7]School of Nursing and Health Studies, University of Dundee, Dundee, UK
[8]Institute of Psychology Health and Society, University of Liverpool, Liverpool, UK
[9]Department of Psychiatry, University of Oxford, Oxford, UK
[10]Primary Care and Population Sciences, Faculty of Medicine, University of Southampton, Southampton, UK
[11]Patient and public involvement contributor, UK
[12]Department of Psychology and Integrated Epidemiology Unit, University of Bristol, Bristol, UK
[13]Department of Sociology, Social Policy and Criminology, University of Liverpool, Liverpool, UK

*Corresponding author larisa.duffy@ucl.ac.uk

**Background:** Despite a growing number of prescriptions for antidepressants (over 70 million in 2018), there is uncertainty about when people with depression might benefit from antidepressant medication and concern that antidepressants are prescribed unnecessarily.

**Objectives:** The main objective of the PANDA (What are the indications for Prescribing ANtiDepressAnts that will lead to a clinical benefit?) research programme was to provide more guidance about when antidepressants are likely to benefit people with depression. We aimed to estimate the minimal clinically important difference for commonly used self-administered scales for depression and anxiety, and to understand more about how patients respond to such assessments. We carried out an observational study of patients with depressive symptoms and a placebo-controlled randomised controlled trial of sertraline

versus placebo to estimate the treatment effect in UK primary care. The hypothesis was that the severity and duration of symptoms were related to treatment response.

**Design:** The programme consisted of three phases. The first phase relied on the secondary analysis of existing data extracted from published trials. The second phase was the PANDA cohort study of patients with depressive symptoms who presented to primary care and were followed up 2, 4 and 6 weeks after a baseline assessment. Both quantitative and qualitative methods were used in the analysis. The third phase was a multicentre randomised placebo-controlled double-blind trial of sertraline versus placebo in patients presenting to primary care with depressive symptoms.

**Setting:** UK primary care in Bristol, London, Liverpool and York.

**Participants:** Patients aged 18–74 years who were experiencing depressive symptoms in primary care. Eligibility for the PANDA randomised controlled trial included that there was uncertainty about the benefits about treatment with an antidepressant.

**Interventions:** In the PANDA randomised controlled trial, patients were individually randomised to 100 mg daily of sertraline or an identical placebo. The PANDA cohort study was an observational study.

**Main outcome measures:** Depressive symptoms measured using the Patient Health Questionnaire were the primary outcome for the randomised controlled trial. Other outcomes included anxiety symptoms using the Generalised Anxiety Disorder-7; depressive symptoms using the Beck Depression Inventory, version 2; health-related quality of life; self-reported improvement; and cost-effectiveness.

**Results:** The secondary analysis of existing randomised controlled trials [GENetic and clinical Predictors Of treatment response in Depression (GenPod), TREAting Depression with physical activity (TREAD) and Clinical effectiveness and cost-effectiveness of cognitive Behavioural Therapy as an adjunct to pharmacotherapy for treatment-resistant depression in primary care (CoBalT)] found evidence that the minimal clinically important difference increased as the initial severity of depressive symptoms rose. Our estimates of minimal clinically important difference were a 17% and 18% reduction in Beck Depression Inventory scores for GenPod and TREAD, respectively. In CoBalT, a 32% reduction corresponded to the minimal clinically important difference but the participants in this study had depression that had not responded to antidepressants. In the PANDA study cohort, and from our analyses in existing data, we found that the minimal clinically important difference varies considerably with the initial severity of depressive and anxiety symptoms. Expressing the minimal clinically important difference as a percentage reduction reduces this variation at higher scores, but at low scores the percentage reduction increased substantially. The results from the qualitative studies pointed out many limitations of the Patient Health Questionnaire-9 items in assessing change and recovery from depression. In the PANDA randomised controlled trial, there was no evidence that sertraline resulted in a reduction in depressive symptoms within 6 weeks of randomisation, but there was some evidence of a reduction by 12 weeks. However, sertraline led to a reduction in anxiety symptoms, an improvement of mental health-related quality of life and an increased likelihood of reporting improvement. The mean Patient Health Questionnaire-9 items score at 6 weeks was 7.98 (standard deviation 5.63) in the sertraline group and 8.76 (standard deviation 5.86) in the placebo group (5% relative reduction, 95% confidence interval –7% to 15%; $p = 0.41$). Of the secondary outcomes, there was strong evidence that sertraline reduced anxiety symptoms (Generalised Anxiety Disorder-7 score reduced by 17% (95% confidence interval 9% to 25%; $p = 0.00005$). Sertraline had a high probability (> 90%) of being cost-effective at 12 weeks. The PANDA randomised controlled trial found no evidence that treatment response or cost-effectiveness was related to severity or duration of depressive symptoms. The minimal clinically important difference estimates suggested that sertraline's effect on anxiety, but not on depression, was likely to be clinically important.

**Limitations:** The results from the randomised controlled trial and the estimates of minimal clinically important difference were not sufficiently precise to provide specific clinical guidance for individuals. We had low power in testing whether or not initial severity and duration of depressive symptoms are related to treatment response.

**Conclusions:** The results of the trial support the use of sertraline and probably other selective serotonin reuptake inhibitors because of their action in reducing anxiety symptoms and the likelihood of longer-term benefit on depressive symptoms. Sertraline could be prescribed for anxiety symptoms that commonly occur with depression and many patients will experience a clinical benefit. The Patient Health Questionnaire-9 items and similar self-administered scales should not be used on their own to assess clinical outcome, but should be supplemented with further clinical assessment.

**Future work:** We need to examine the longer-term effects of antidepressant treatment. We need more precise estimates of the treatment effects and minimal clinically important difference at different severities to provide more specific guidance for individuals. However, the methods we have developed provide an approach towards providing such detailed guidance.

**Trial registration:** Current Controlled Trials ISRCTN84544741 and EudraCT number 2013-003440-22.

**Funding:** This project was funded by the National Institute for Health Research (NIHR) Programme Grants for Applied Research programme and will be published in full in *Programme Grants for Applied Research*; Vol. 7, No. 10. See the NIHR Journals Library website for further project information.

# Contents

# List of figures

# List of abbreviations

| | |
|---|---|
| BDI | Beck Depression Inventory |
| BDI-II | Beck Depression Inventory, version 2 |
| CI | confidence interval |
| CIS-R | Clinical Interview Schedule – Revised |
| CoBalT | Clinical effectiveness and cost-effectiveness of cognitive Behavioural Therapy as an adjunct to pharmacotherapy for treatment-resistant depression in primary care |
| CRN | Clinical Research Network |
| DSM-IV | *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition |
| EQ-5D | EuroQol-5 Dimensions |
| EQ-5D-5L | EuroQol-5 Dimensions, five-level version |
| EVPPI | expected value of partial perfect information |
| GAD-7 | Generalised Anxiety Disorder-7 |
| GenPod | GENetic and clinical Predictors Of treatment response in Depression |
| GP | general practitioner |
| GRC | Global Rating of Change |
| HAMD | Hamilton Rating Scale for Depression |
| HAMD-17 | Hamilton Rating Scale for Depression-17 items |
| HAMD-24 | Hamilton Rating Scale for Depression-24 items |
| ICD-10 | *International Classification of Diseases*, Tenth Edition |
| MADRS | Montgomery–Åsberg Depression Rating Scale |
| MCID | minimal clinically important difference |
| NICE | National Institute for Health and Care Excellence |
| NRES | National Research Ethics Service |
| PANDA | What are the indications for Prescribing ANtiDepressAnts that will lead to a clinical benefit? |
| PHQ-9 | Patient Health Questionnaire-9 items |
| PPI | patient and public involvement |
| QALY | quality-adjusted life-year |
| RCT | randomised controlled trial |
| ROC | receiver operator characteristic |
| SD | standard deviation |
| SF-36 | Short Form questionnaire-36 items |
| SSRI | selective serotonin reuptake inhibitor |
| SURF | Service Users Research Forum |
| TREAD | TREAting Depression with physical activity |

# Plain English summary

There were over 70 million antidepressant prescriptions in England in 2018, and there are concerns that they are overprescribed. The aim of the PANDA (What are the indications for Prescribing ANtiDepressAnts that will lead to a clinical benefit?) research programme was to provide general practitioners with improved guidance to help them make recommendations about the likely response to antidepressants.

We carried out two studies. In the first, we recruited 558 people with depressive symptoms and followed them for 6 weeks. We estimated a patient-centred measure of the minimal clinically important difference, an improvement that would be recognised by the patient. We found that the minimal clinically important difference was best expressed as about a 20% reduction in the patient's initial symptoms, but at lower symptoms the percentage change for the minimal clinically important difference became larger.

We also interviewed patients with open-ended interviews. They reported that the Patient Health Questionnaire-9 items failed to fully capture their experience of recovery from depression. Some patients struggled with how the questions were phrased and there was a lot of disagreement between self-reported improvement and change in Patient Health Questionnaire-9 items score. We concluded that clinicians should not rely on scales such as the Patient Health Questionnaire-9 items to assess improvement but should use additional questions and further clinical assessment before deciding on improvement.

Our second study was a randomised clinical trial (on 653 participants) to investigate the clinical effectiveness of sertraline, a commonly used first-line antidepressant. We found no evidence that sertraline was more effective than a placebo (an identical inactive capsule) at reducing depression symptoms by 6 weeks, but by 12 weeks we found some evidence for a reduction in depressive symptoms. However, we found strong evidence that sertraline was effective at reducing anxiety symptoms and that patients who took sertraline were more likely to report improvement and better mental health overall. Sertraline is a cheap intervention that had a high probability of being cost-effective. We found no evidence that the severity and duration of patients' depressive symptoms predicted their response to antidepressants or the cost-effectiveness of treatment.

# Scientific summary

## Background

There were over 70 million antidepressent prescriptions in England in 2018, with a substantial cost to the NHS. Selective serotonin reuptake inhibitors are the first-line antidepressant recommended by UK National Institute for Health and Care Excellence guidelines. However, there is little clinical guidance on when an antidepressant should be prescribed, uncertainty about when patients might benefit and concerns that antidepressants are prescribed unnecessarily.

## Aims and objectives

The overall aim of the PANDA (What are the indications for Prescribing ANtiDepressAnts that will lead to a clinical benefit?) research programme was to provide general practitioners with improved guidance that would enable them to make recommendations about the probable response to antidepressants for patients with depressive symptoms. We wanted to estimate the minimal clinically important difference for depressive symptom questionnaires and understand more about how patients respond to such assessments. We carried out a randomised controlled trial (the PANDA randomised controlled trial) that examined the cost-effectiveness of sertraline compared with placebo. Our hypothesis was that the response to antidepressants compared with placebo would increase with both the severity and duration of depressive symptoms. We conducted our research in three phases.

## Methods and results

### *Phase 1: using previously collected data*

#### Aim 1a: using existing data to estimate the minimal clinically important difference for the Beck Depression Inventory, version 2

We examined data from three existing randomised controlled trials [GENetic and clinical Predictors Of treatment response in Depression (GenPod), TREAting Depression with physical activity (TREAD) and Clinical effectiveness and cost-effectiveness of cognitive Behavioural Therapy as an adjunct to pharmacotherapy for treatment-resistant depression in primary care (CoBalT)], which used a Global Rating of Change question and the Beck Depression Inventory, version 2, to measure depressive symptoms. We estimated the minimal clinically important difference by calculating the reduction in Beck Depression Inventory, version 2, scores in those who reported they had improved. We found evidence that the minimal clinically important difference increased as the initial severity of depressive symptoms rose and the minimal clinically important difference was better described as a percentage reduction of the initial score rather than an absolute fixed value. Our estimates of minimal clinically important difference were a 17%, 18% and 32% reduction for people with depression for GenPod, TREAD and CoBalT, respectively.

#### Aim 1b: 'mapping' the relationship between different depression scales

We identified 31 RCTs that had included more than one depressive symptom, or health-related quality-of-life scales. We used a novel method to compare the relative responsiveness of the scales that allowed estimation of mapping coefficients that could translate treatment effects. We found evidence that, of the depression measures, the Patient Health Questionnaire-9 items was the most responsive to change. A 1.0 standard deviation treatment effect on the Beck Depression Inventory (the reference) was, on average, equivalent to 1.52 standard deviations on the Patient Health Questionnaire-9 items (95% credibility interval 1.17 to 2.05 standard deviations) and 1.31 standard deviations on the Hamilton Rating Scale for Depression (95% credibility interval 1.04 to 1.69 standard deviations).

### Aim 1c: value-of-information study to estimate the probable benefit of carrying out the PANDA randomised controlled trial

We developed a novel economic model that incorporates the severity of depression as part of the decision-making process. The model determined that treating patients with a severity score of $\geq 2$ on the Hamilton Rating Scale for Depression had the highest probability (> 65%) of being cost-effective at a £20,000 willingness-to-pay threshold. However, there was a lack of evidence at low levels of severity, and the results relied on a number of assumptions. We concluded that the PANDA randomised controlled trial was likely to be an efficient use of resources to reduce uncertainty in the most cost-effective treatment for such patients (expected value of partial perfect information = £67.7M over a 10-year time horizon).

## *Phase 2: cohort study: using both quantitative and qualitative methods*

We conducted a cohort study of 558 people who had presented with symptoms of depression or low mood to their general practitioner in the previous 12 months. Practices were recruited in Bristol, Liverpool and York. Potential participants were identified by searching the general practitioner electronic records and were then invited to participate if they were aged 18–70 years and did not have comorbid psychosis, bipolar disorder, eating disorders or substance dependence. Participants were followed up at 2, 4 and 6 weeks after baseline, when they completed self-administered questionnaires assessing symptoms of depression and anxiety (i.e. Beck Depression Inventory, version 2, Patient Health Questionnaire-9 items and Generalised Anxiety Disorder-7) and were asked to rate their own improvement using a Global Rating of Change question. The Global Rating of Change was assessed by asking patients '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), or 'I feel a lot worse' (5). The Clinical Interview Schedule – Revised was completed at baseline.

### Aim 2a: estimates of the minimal clinically important difference in commonly used self-administered questionnaires for depressive symptoms

In the PANDA cohort, we estimated the optimal threshold score for the Patient Health Questionnaire-9 items, Beck Depression Inventory, version 2, and Generalised Anxiety Disorder-7 (a measure of anxiety symptoms), below which someone was more likely to report feeling better. This as the most robust estimate of minimal clinically important difference as it also takes account of the variability of scores. The PANDA cohort had a lower range of scores than those found in the previous analysis (see *Aim 1a: using existing data to estimate the minimal clinically important difference for the Beck Depression Inventory, version 2*) and we found that the size of the minimal clinically important difference, expressed as proportional reduction, was larger for when initial severity of depressive symptoms was low. For those with an initial score of 12 on the Patient Health Questionnaire-9 items, the minimal clinically important difference was 19.7%, but for those with an initial score of 4, the minimal clinically important difference was 48.2%.

### Aim 2b: investigation of the changes reported by patients as they recover from depression

#### *Study 1: usefulness of the Patient Health Questionnaire-9 items in primary care to determine meaningful symptoms of low mood – a qualitative study*

We conducted a longitudinal qualitative study of 18 participants selected using the same criteria as the PANDA cohort but with a purposive sampling strategy. The participants completed the Global Rating of Change and the Patient Health Questionnaire-9 items using cognitive interviewing techniques at 2, 4 and 6 weeks after baseline. Participants reported that the Patient Health Questionnaire-9 items omits certain symptoms that are important to people with depression. Participants translated the options on frequency such that 'several days' was used to represent a higher intensity of symptom. Participants thought that the Global Rating of Change was a good way of taking account of all the symptoms and changes that were important to them.

### *Study 2: variation in recognition of happy and sad facial expressions and self-reported depressive symptom severity*

Biases in the way that people process emotional information might be markers of recovery from depression and it has been suggested that antidepressants work via their action on emotion-processing. In the PANDA cohort, we administered a task concerned with identifying the emotional expression on morphed faces. We found that depressive symptoms were associated with an increase in reporting happy faces when ambiguous expressions were presented. There was no association between depressive symptoms and recognition of sad faces.

### *Study 3: variation in the recall of socially rewarding information and depressive symptom severity – a prospective cohort study*

We investigated whether or not severity of depressive symptoms in the PANDA cohort was associated with recall of positive and negative words, as a measure of emotional processing. We found evidence that, for every increase in two positive words recalled, depressive symptoms were lower by –0.6 (95% confidence interval –1.0 to –0.2) Beck Depression Inventory points, but there was no evidence for an association with negative words. These findings suggest that people with more severe depressive symptoms recall less positive information but negative information is unaffected, which is similar to the face recognition findings above.

## Aim 2c: to investigate disagreement between self-reported improvement and changes in the scores on depressive symptom questionnaires

### *Study 1: why are there discrepancies between depressed patients' Global Rating of Change and scores on the Patient Health Questionnaire-9 items depression module? A qualitative study of primary care in England*

We identified participants from the PANDA cohort who reported a disagreement between the Global Rating of Change and the change in scores on the Patient Health Questionnaire-9 items. We defined disagreement as meaningful if scores changed by ≥ 15%, based on preliminary results. Of the first 86 participants from the Liverpool site, 29 participants (34%) with the most pronounced disagreement took part in this qualitative study. We identified four themes used by participants to explain the mismatch between the Global Rating of Change and the Patient Health Questionnaire-9 items: (1) problems with the Patient Health Questionnaire-9 items and perceptions that the Global Rating of Change provided a more accurate assessment of current mental state, (2) the impact of recent positive or negative life events, (3) personal understanding of depression and coping mechanisms and (4) an inability to recall how they felt in the past.

### *Study 2: changes in self-administered measures of depression severity and patients' own perceptions of changes in their mood – a cohort study in primary care*

This quantitative study in the PANDA cohort investigated the disagreement between changes in Patient Health Questionnaire-9 items and Beck Depression Inventory, version 2, scores and responses to self-reported improvement (Global Rating of Change). We used a minimal clinically important difference estimate of 20%. We found that in a substantial proportion of patients (51% on the Patient Health Questionnaire-9 items scale and 55% on the Beck Depression Inventory, version 2, scale), there was a clinically important disagreement between their responses to the questionnaires and the Global Rating of Change. We found that participants who reported anxiety and poor quality of life were less likely to report feeling better on the Global Rating of Change after taking account of their change in depressive symptom score.

## *Phase 3: randomised controlled trial*

## Aim 3: severity and duration of depressive symptoms and response to antidepressants (the PANDA trial) – a pragmatic randomised controlled trial in primary care

We conducted a pragmatic randomised multicentre double-blind placebo-controlled trial of patients from 179 primary care surgeries in four UK sites. Patients aged 18 to 74 years with reported depressive symptoms were eligible if there was uncertainty about the benefit of an antidepressant. Patients were

individually randomised to 100 mg daily of sertraline or placebo. The primary outcome was the Patient Health Questionnaire-9 items score at 6 weeks. Secondary outcomes at 2, 6 and 12 weeks were depressive symptoms and remission assessed using the Patient Health Questionnaire-9 items and the Beck Depression Inventory, version 2, generalised anxiety disorder, mental and physical health-related quality of life, global ratings of change, health-care costs and quality-adjusted life-years.

Our primary outcome analyses were of 550 patients (sertraline, $n = 266$; placebo, $n = 284$). The mean Patient Health Questionnaire-9 items score at 6 weeks was 7.98 (standard deviation 5.63) in the sertraline group and 8.76 (standard deviation 5.86) in the placebo group. In the sertraline group, Patient Health Questionnaire-9 items scores were 5% (95% confidence interval –7% to 15%; $p = 0.41$) lower than in the placebo group. There was no evidence that the treatment effect on the primary outcome varied according to depression severity or duration. Of the secondary outcomes, there was strong evidence that sertraline reduced anxiety symptoms (Generalised Anxiety Disorder-7 score reduced by 17%, 95% confidence interval 9% to 25%; $p = 0.00005$), improved mental, but not physical, health-related quality of life and self-reported global improvement. There was weak evidence that depressive symptoms were reduced by sertraline at 12 weeks for both the Patient Health Questionnaire-9 items and the Beck Depression Inventory, version 2, scales. There was some evidence that sertraline was more cost-effective than the placebo at a threshold of £20,000 (incremental net monetary benefit £118, 95% confidence interval –£23 to £260) per quality-adjusted life-year. Sertraline had a high probability (> 90%) of being cost-effective if the health system was willing to pay > £20,000 per quality-adjusted life-year gained. We did not find evidence for any influence of severity of symptoms or duration on treatment response or cost-effectiveness, but our analysis had low statistical power.

## Conclusions

### *Use of self-administered questionnaires in assessing depressive symptoms*

There is a strong correlation between changes in Patient Health Questionnaire-9 items score and self-reported improvement using the Global Rating of Change, and the Patient Health Questionnaire-9 items appeared to be better at detecting average change than the Beck Depression Inventory, version 2, scale and the Hamilton Rating Scale for Depression. However, we identified substantial disagreement between the Global Rating of Change and changes in Patient Health Questionnaire-9 items score when considering individual responses. Up to half of the people in our sample reported a disagreement between the Global Rating of Change and changes on the Patient Health Questionnaire-9 items. Our qualitative research also pointed out the limitation of these scales and supported the validity of the Global Rating of Change in providing an overall measure of improvement. We conclude that the Patient Health Questionnaire-9 items should not be used on its own to assess individual change. It is important for clinicians to supplement results from the Patient Health Questionnaire-9 items with additional clinical assessment including open-ended questions about any improvements.

We have estimated the minimal clinically important difference for the Beck Depression Inventory, version 2, Patient Health Questionnaire-9 items and Generalised Anxiety Disorder-7 using a patient anchoring approach in which we use the Global Rating of Change to estimate the threshold below which someone is more likely to regard themselves as having improved. At higher initial severity, we found a minimal clinically important difference of about 20%. However, at lower severities, the minimal clinically important difference increased as a proportion; so for the lowest severity we examined, the minimal clinically important difference was as large as 50% for the Patient Health Questionnaire-9 items. Our estimates were also imprecise at the lower severities. This complex relationship between initial severity and minimal clinically important difference is an important finding.

### *Effectiveness and cost-effectiveness of sertraline*

The PANDA randomised controlled trial did not find evidence that sertraline led to a clinically important reduction in depressive symptoms at 6 weeks. However, we found strong evidence at 6 and 12 weeks that sertraline reduced anxiety symptoms, that mental health-related quality of life improved and that participants reported feeling better. There was some weak evidence that sertraline reduced depressive symptoms at 12 weeks. Sertraline is an inexpensive intervention that had a high probability (> 90%) of being cost-effective compared with placebo at 12 weeks. Our results of an improvement in anxiety symptoms are to be expected given previous findings, but the lack of an early effect on depressive symptoms is unexpected. We think that the most likely explanation for this is that previous trials have mostly used the Hamilton Rating Scale for Depression or similar measures of depression. The Hamilton Rating Scale for Depression is a relatively unstandardised measure in which the observer can alter questions and it requires judgements; therefore, this might have led to a halo effect on the rating of depressive symptoms because participants reported feeling better as anxiety symptoms had reduced. We could not find evidence that the treatment response to sertraline or cost-effectiveness varied according to severity or duration, but these analyses had low power.

We can apply our minimal clinically important difference results to the PANDA randomised controlled trial. At 6 weeks' follow-up, our findings suggest that there was a 5% (95% confidence interval –7% to 15%) reduction in Patient Health Questionnaire-9 items scores and a 21% (95% confidence interval 11% to 30%) reduction in Generalised Anxiety Disorder-7 scores. The relevant minimal clinically important difference for Patient Health Questionnaire-9 items is 21% and for Generalised Anxiety Disorder-7 is 27%. We can therefore conclude that, on average, sertraline does not lead to a clinically important difference in depressive symptoms at 6 weeks, but the effect on anxiety symptoms is more consistent with such a clinically important effect. A further implication from our minimal clinically important difference results is that, for those who expect to have low scores at follow-up, the minimal clinically important difference is very large. About 30% of the PANDA randomised controlled trial placebo group had a Generalised Anxiety Disorder-7 score of $\leq 3$ at follow-up, which suggests that they would not have benefited from sertraline if they had received that treatment.

An alternative approach would be to estimate the proportion who reported 'feeling better' using our self-reported improvement question. This allows estimation of the number needed to treat. The number needed to treat at 6 weeks was 8.5 (95% confidence interval 5.2 to 22.1) participants and at 12 weeks 6.4 (95% confidence interval 4.6 to 10.3) participants.

## Trial registration

This trial is registered as ISRCTN84544741 and EudraCT number 2013-003440-22.

## Funding

This project was funded by the National Institute for Health Research (NIHR) Programme Grants for Applied Research programme and will be published in full in *Programme Grants for Applied Research*; Vol. 7, No. 10. See the NIHR Journals Library website for further project information.

# SYNOPSIS

## Background

Depression is the leading cause of disability globally.[1] Most people with depression or depressive symptoms are treated in primary care and antidepressants are often the first-line treatment. There were over 70 million antidepressent prescriptions in England in 2018.[2] There is much uncertainty about when people with depression might benefit from an antidepressant and concern that antidepressants are overprescribed. General practitioners (GPs) often feel under pressure to provide some treatment and have to make a difficult decision about whether or not an individual will benefit from an antidepressant. The existing guidelines use terms such as 'mild' or 'moderate' depression to guide prescription but do not specify what this means, and such guidelines are not based on empirical studies. Identifying the patients who are most likely to respond to an antidepressant would improve the management of depression in primary care as well as in other clinical settings. This would reduce the inappropriate prescription of antidepressants in those less likely to respond, and increase appropriate prescription in those more likely to respond.

## The minimal clinically important difference

To make informed recommendations about when treatments are of benefit to patients, we must decide what constitutes a clinically important treatment effect. There is no consensus about the size of a 'clinically important difference' on continuous outcome scales. The National Institute for Health and Care Excellence (NICE) guideline group, which includes service users, suggested that this difference was three points on the Hamilton Rating Scale for Depression (HAMD),[3] but did not provide any empirical justification.[4] We need to decide on a clinically important treatment effect before we can make recommendations about when selective serotonin reuptake inhibitors (SSRIs) are of benefit and to estimate the sample size of a study. There have been previous attempts to determine clinically important differences[5,6] but these have relied on classifying people as 'well' or 'ill' and then calculating the differences in score between the groups. However, these metrics do not take into account the perspectives of the patient. We were interested in participants' own views about improvement to determine a clinically important difference. Our approach was to ask people to rate their own improvement and then use this to calculate the difference in scores corresponding to a change. This method has been used in relation to quality-of-life scales but, as far as we are aware, not in relation to depression.[7] There has been work comparing the results of self-administered questionnaires with psychiatric diagnostic assessments.[8] Our approach was therefore in contrast to these methods and emphasises the patient's perspective.

## Measurement of depression

Another barrier to the development of evidence-based guidelines for antidepressant prescription in primary care is the lack of a standardised measure of depressive symptoms that is easily implemented. Existing studies on antidepressants mostly use rating scales, such as HAMD, that are difficult to standardise, are designed for clinician administration and require training. It is extremely unlikely that these scales would ever be used in primary care and this has never been proposed, to our knowledge. Shorter self-administered scales, such as the Patient Health Questionnaire-9 items (PHQ-9),[8] have been used in UK primary care and in the NHS Improving Access to Psychological Treatment services. However, it is widely thought that the PHQ-9 does not provide sufficiently good data to guide prescribing even if it is useful as an outcome measure. The NICE CG90 depression guideline[9] explicitly recommends not using the PHQ-9 and similar scales alone for the purpose of guiding treatment.

Short questionnaires, such as the PHQ-9, are unlikely to give accurate information to guide prescription, but evidence suggests that they do perform well as a measure of outcome or change.[10] However, this symptom-based approach to measuring outcome has been challenged by the growing literature about 'recovery', largely from the perspective of people with psychosis. Two areas highlighted in this literature include the concepts of hope and empowerment, the notion that the patient feels able to change.[11] We are aware of only a limited literature concerned with 'recovery' in people who have experienced depression.[12] Malpass *et al.*[13] have also interviewed people recovering from depression in relation to change on the PHQ-9 and did identify some areas that were not well covered in that questionnaire. These included anxiety along with 'awareness' and 'ability to make changes'.

There is still some uncertainty about how antidepressants work but Harmer *et al.*[14] have found that antidepressants lead to changes early on in the processing of emotional information. Beck *et al.*[15] first proposed that negative interpretations, beliefs and memories play a key role in depression and developed cognitive–behavioural therapy. Subsequently, the cognitive neuropsychological model of depression has proposed that lower-level changes in emotional processing play a causal role in the genesis of symptoms and precede changes in depressive symptoms.[14] Simple tests, such as face recognition and word recall, could be used alongside symptom measures to investigate changes in depression and complement measures of symptoms. Existing studies have been small and used case–control designs that are prone to selection bias. If we could confirm that emotion-processing tasks were related to depressive symptoms, future research could see if the response to emotion-processing tasks could be used to predict clinical response. We therefore also investigated associations between depressive symptom severity and emotional processing and how this might change in recovery.

There are a wide range of existing outcome assessments of depressive symptoms, for example HAMD, PHQ-9, Beck Depression Inventory, version 2 (BDI-II)[16] and the Hospital Anxiety and Depression Scale (HADS).[17] It would be useful to know how these scales inter-relate so that the results of clinical trials can be compared with each other.[18,19] If we could 'map' the scores on the scales against each other, this would help with the interpretation of existing data and in the application of existing and future results to clinical practice.

## What factors are associated with response to antidepressants?

It has been proposed that antidepressants are more effective for patients with more severe depression, but the evidence for this is inconsistent[20–25] and recent large studies of individual patient data suggest no influence of depression severity.[23–25] On the other hand, a systematic review of patients with depressive symptoms not meeting diagnostic criteria[26] did not find evidence that antidepressants were effective. However, the majority of the existing trials were not designed to investigate this hypothesis and exclude patients who are below a certain severity threshold.[27,28] A more general criticism is that the current evidence is dominated by trials performed for regulatory purposes so it is difficult to generalise their results to patients currently receiving treatment, mainly provided in primary care.

To help GPs decide whether or not to prescribe SSRIs, it is important to compare a SSRI with a placebo. This becomes more rather than less important in less severe depression as the differences between placebo and active medication are likely to be smaller than for more severe depression.[20] By 'response' to antidepressants, we refer to the difference between the antidepressant and the placebo.

The two main diagnostic manuals, *International Classification of Diseases*, Tenth Edition (ICD-10),[29] and *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition (DSM-IV),[30] have similar, but not identical, diagnostic criteria for a depressive episode (ICD-10) and major depression (DSM-IV). DSM-IV also has a category of minor depression that requires fewer symptoms. As a rule, GPs do not use these

diagnostic criteria and there is currently little empirical evidence that meeting either of these diagnostic criteria can alone indicate whether or not antidepressants will be beneficial. There is a consensus that depressive symptoms can be viewed along a continuum of severity, and, in population terms, 'subthreshold' symptoms are still a public health concern.[31,32] It is therefore important to assess the severity of symptoms along this continuum. People who do not meet diagnostic criteria might still benefit from antidepressant treatment, or the converse.

There is evidence that antidepressants can be effective for people with dysthymia.[33,34] Dysthymia is a US term used in DSM-IV but not in ICD-10. It describes depressive symptoms of long duration ($\geq$ 2 years) but not meeting the diagnostic criteria for depression. NICE guidelines[19] recommend SSRIs for persistent subthreshold depressive symptoms but give no guidance on the duration of the persistence. We proposed, as adopted by current UK guidelines,[19,35] that symptom severity and duration of symptoms are two separate dimensions that both might help to predict response to antidepressants.

## The PANDA research programme

The overall aim of the PANDA (What are the indications for Prescribing ANtiDepressAnts that will lead to a clinical benefit?) research programme was to provide GPs or other clinicians with improved guidance about when antidepressants are most likely to result in a clinical benefit for patients with depressive symptoms in primary care and be cost-effective for the NHS. The main hypothesis was that response to antidepressants (compared with placebo) would increase with both the severity and duration of depressive symptoms. The programme aimed to investigate whether or not there are thresholds of severity and duration above which a clinically important response to antidepressants is most likely. In line with growing emphasis on patient-centred care, 'clinically important' was defined as a reduction in depressive symptoms large enough that patients would detect feeling better. An additional aim was therefore to establish, for the first time, the reduction in depressive symptoms on self-administered depression questionnaires that is required for patients to feel better. To assess severity and duration, we used a detailed, self-administered, computerised assessment that could be easily implemented in primary care. Finally, we also wanted to investigate how patients interpreted the questions in the current self-administered scales, to investigate how that related to self-reported improvement and to investigate the relationship between depressive symptoms and the underlying cognitive biases that are characteristic of depression.

A new randomised controlled trial (RCT) was required to provide this information, despite the wealth of data available from previous placebo-controlled studies that have largely been carried out by the pharmaceutical industry and are comprehensively summarised in Cipriani *et al.*[28] First, the existing data were mostly of a poor quality and were carried out decades ago for regulatory purposes rather than to study clinical indications. Cipriani *et al.*[28] noted that 78% of trials were industry funded and that, overall, trials were of poor quality and small, with a mean sample size of 224 (across arms). In addition, 82% of trials were also at moderate or high risk of bias and the larger more recent placebo-controlled trials reported smaller effect sizes, perhaps reflecting more rigorous methods. Second, existing results on the influence of depression severity on antidepressant response are in terms of HAMD scores. This is not useful in primary care because of the training required. Even though we should be able to calculate equivalent scores on measures such as the PHQ-9, more detailed assessments would provide a better predictor of response than brief questionnaires. Third, existing data are unlikely to generalise to the population currently receiving antidepressants in primary care in the UK or in other countries. Most of the trials excluded people at lower levels of severity and the recruitment methods are usually unknown. Finally, we are not aware of any existing data from previous studies on the influence of duration of illness.

## The objectives of the PANDA research programme

The PANDA research programme consisted of three phases; the aims of each phase and the title of corresponding PANDA papers are listed below.

### *Phase 1: using previously collected data*

Aim 1a: to use existing data to estimate the minimal clinically important difference (MCID) for the BDI-II.

- See *Appendix 1*: Button *et al.*[36]

Aim 1b: to 'map' the relationship between different depression scales to estimate the scores on each scale that correspond to the same severity of symptoms.

- See *Appendix 2*: Kounali *et al.*[37]

Aim 1c: to carry out a value-of-information study to estimate the probable benefit of carrying out a RCT as described in phase 3.

- See *Appendix 3*: Thom *et al.*[38]

### *Phase 2: cohort study – using both quantitative and qualitative methods*

Aim 2a: to estimate the MCID in commonly used self-administered questionnaires for depressive symptoms.

- See *Appendix 4*

Aim 2b: to investigate the changes reported by patients as they recover from depression.

- See *Appendix 5*: Malpass *et al.*[39]
- See *Appendix 6*: Bone *et al.*[40]
- See *Appendix 7*: Lewis *et al.*[41]

Aim 2c: to investigate disagreement between self-reported improvement and changes in the scores on depressive symptom questionnaires.

- See *Appendix 8*: Robinson *et al.*[42]
- See *Appendix 9*

### *Phase 3: randomised controlled trial*

Aim 3: to investigate the severity and duration of depressive symptoms that are associated with a clinically important response to sertraline in people with depression and whether or not these factors are associated with the cost-effectiveness of sertraline.

- See *Appendix 10*: Salaminios *et al.*[43]
- See *Appendix 11*: Lewis *et al.*[44]
- See *Appendix 12*: Hollingworth *et al.*[45]

The inter-relationships between the three phases are summarised in *Figure 1*.

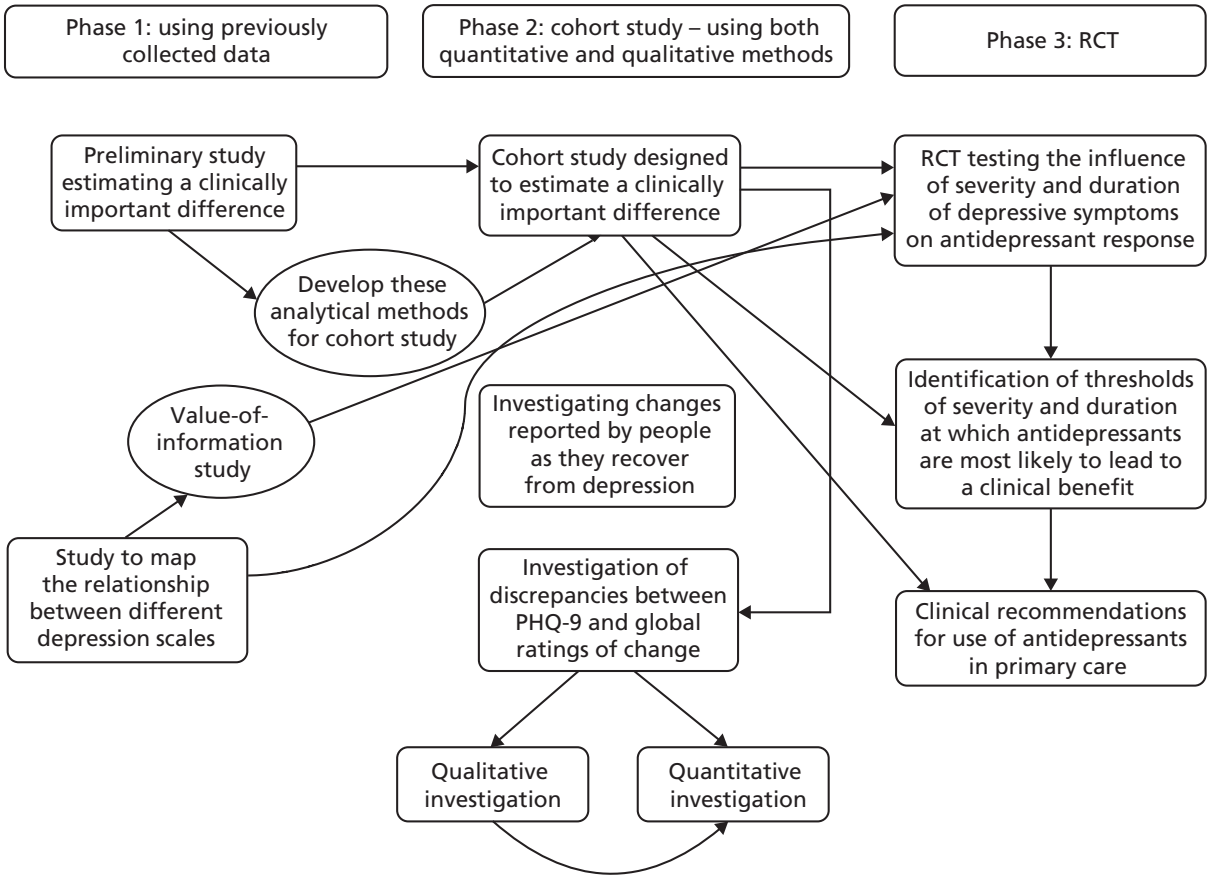



**FIGURE 1** The inter-relationships between the three phases.

## Changes to the programme

1. In our research proposal, we originally had aimed to carry out a systematic review and individual patient data meta-analysis to investigate the relationship between the severity of depressive symptoms and the response to antidepressants. However, after the research programme was funded, an individual patient data meta-analysis was published by Gibbons *et al.*[23] using data provided by the pharmaceutical industry and there have been more recent replications.[24,25] Gibbons *et al.*[23] analysed all the placebo-controlled studies of the antidepressants fluoxetine and venlafaxine that were sponsored by the relevant pharmaceutical company and had therefore obtained a complete and unbiased sample of the placebo-controlled trials. As a result, we decided to abandon the original proposal for a systematic review in this area based on literature searching and approaching authors for individual patient data.
2. With the resources that had originally been earmarked for the individual patient data meta-analysis, we carried out an analysis of existing data to begin our study of the MCID (aim 1a and see *Appendix 1*) and a value-of-information study (aim 1c and see *Appendix 3*). Previously published economic models including that used in the NICE CG90[19] did not consider treatment by baseline severity and none modelled depression severity itself, so this required the development of an economic model in which severity of depression was part of the decision-making process.
3. We also included some simple emotion-processing tasks in the PANDA cohort study. These tasks are influenced by antidepressant medication and might indicate changes that occur during recovery from depression that are not currently assessed by self-administered questionnaires. These tasks therefore extended aim 2b and the relationship with depressive symptoms and changes in symptoms are reported in *Appendices 6* and *7*.

4. When originally devised, the primary aim of the RCT was to investigate the severity and duration of symptoms associated with a clinically important response to sertraline, as stated in the protocol paper[43] and on the International Standard Randomised Controlled Trial Number (ISRCTN) registry. However, it was apparent towards the later stages of designing the RCT and in formulating the detailed analysis plan[46] (uploaded before any analyses were performed to http://discovery.ucl.ac.uk and approved by the Trial Steering Committee) that we would have insufficient statistical power to estimate plausible interaction effects that would allow us to investigate those aims. Our power calculation and primary analysis (as stated in the analysis plan[46]) are therefore based on a primary aim to examine the clinical effectiveness of sertraline versus placebo. Interactions between severity and duration at baseline and treatment response were planned as exploratory.
5. In the original funding application, we used citalopram as our choice of antidepressant. However, during the set-up stage, new guidance was released informing clinicians that citalopram can prolong the QT interval, especially in higher doses. Although the risk was low, we decided to change the study medication to sertraline, another commonly prescribed SSRI that is no longer under patent. There are very few pharmacological differences between the SSRIs and we believe that the results of our study can be applied to all SSRIs.

# Phase 1: using previously collected data

## Aim 1a: to use existing data to estimate minimal clinically important difference from the patient perspective

### *Research aims*

The aim of this study was to use existing data to estimate the MCID from the patient's perspective, and to investigate whether or not the MCID varied according to how severely ill patients were to begin with, which we have called 'baseline dependency'.

### *Methods for data collection*

We used existing data from three RCTs [GENetic and clinical Predictors Of treatment response in Depression (GenPod), TREAting Depression with physical activity (TREAD) and Clinical effectiveness and cost-effectiveness of cognitive Behavioural Therapy as an adjunct to pharmacotherapy for treatment-resistant depression in primary care (CoBalT)],[47–49] in which we had asked patients a Global Rating of Change (GRC) question. All these studies recruited individuals who met the ICD-10 criteria for depression. GENPOD compared citalopram and reboxetine, TREAD evaluated an intervention designed to increase physical activity and CoBalt investigated cognitive–behavioural therapy as an adjunctive treatment in people who had not responded to antidepressants. As far as we are aware, these are the only trials that used the GRC. Each trial used the BDI-II as the primary outcome, which provided an opportunity to estimate the MCID for the BDI-II. Each RCT investigated treatment options for depression and followed participants over several months providing at least two time periods for analysis. Data from 1039 patients who met the ICD-10 diagnostic criteria for depression were analysed.

### *Analysis*

To test whether or not the MCID varied according to baseline severity, we assessed change in BDI-II scores as both absolute (difference) and proportional reduction. Participants were dichotomised into 'better' and 'not better' (combining feeling the same and feeling worse) using the GRC. To examine whether or not the MCID varied according to baseline severity of depression, and thus determine whether or not MCID was best assessed in terms of absolute change or per cent reduction in scores from baseline, we used generalised linear models. We used receiver operator characteristic (ROC) analysis to find the change in BDI-II score (the 'cut-off point' or threshold) that optimally classifies those individuals who felt better and those who did not.

### *Key findings*

We found strong evidence that the size of the MCID depended on the initial severity of depressive symptoms. Patients with more severe depressive symptoms at baseline required a larger change in their BDI-II scores to report feeling better. Participants in the CoBalt study whose symptoms had not responded to antidepressants needed to experience larger changes on the BDI-II (on average) to report feeling better. Overall, for every 10-point increase in baseline severity on the BDI-II, the mean score associated with feeling better increased by 4.8 points [95% confidence interval (CI) 0.9 to 8.5 points] on the BDI-II. There was statistical evidence that the MICD was best assessed as a proportional reduction of scores rather than an absolute fixed value.

Our best estimate for the MCID based on the ROC analyses provide the best estimates of MCID, with an improvement from baseline of 17%, 18% and 32% for GenPod, TREAD and CoBalT, respectively. As noted above, the MCID for CoBalt, in which the participants had depression that had not responded to antidepressants, was larger than for the other two studies.

### *Limitations*

The study was a secondary data analysis from existing trials and the use of RCT data introduced regression to the mean and this complicated the interpretation of the changes. The trials included only those who met the ICD-10 criteria for depression so there were fewer data at lower levels of severity where there is more controversy about the MCID.

### *Inter-relation with other parts of the programme*

This study allowed us to develop our approach towards estimating the MCID in preparation for the PANDA cohort study in phase 2. The MCID was also used to inform our power calculation for the RCT in phase 3. The MCID estimate was also used for aim 2c, in which we investigated disagreements between self-reported improvement and the changes in self-administered depression scales.

## Aim 1b: 'mapping' the relationship between different depression scales

### *Research aims*

The aim was to estimate the relative responsiveness of commonly used scales for depressive symptoms. This allows comparison of treatment effects across different studies and also allowed us to draw conclusions about the relative responsiveness of outcome measures that might not have been directly compared.

### *Methods for data collection*

A search for all measures of depression, anxiety and quality-of-life outcomes was conducted in May 2011 in studies on the Cochrane Depression Anxiety and Neurosis review group's register. We identified 31 placebo and usual-care controlled studies with clearly defined treatment and control groups that reported two or more outcome measures of depression or quality of life. Eleven of the studies were drug trials and the remaining studies were of psychological therapies. The depression measures included were the Beck Depression Inventory (BDI), PHQ-9, Hamilton Rating Scale for Depression-17 items (HAMD-17),[50] Hamilton Rating Scale for Depression-24 items (HAMD-24)[50] and Montgomery–Åsberg Depression Rating Scale (MADRS).[51] We also examined the EuroQol-5 Dimensions, five-level version (EQ-5D-5L),[52] Short Form questionnaire-36 items (SF-36) mental capacity score and SF-36 physical capacity score.[53]

### *Analysis*

We used a new meta-analytic method that had been developed by our co-investigators[54] that can be interpreted as estimating the relative responsiveness of different scales. The data used in the study were the mean treatment differences between active and control arms after 12 weeks' follow-up or as close as possible to that. If this was unavailable, we used the difference between the mean score at baseline and follow-up. The pooled standard deviation (SD) at follow-up was used for standardisation, if available. If this was not available, the pooled SD on the difference between the mean score at baseline and follow-up was used. The analysis used methods that allowed for simultaneous estimation of treatment effects on continuous outcomes and the 'mappings' between treatment effects in a Bayesian framework. The mappings are ratios of the underlying treatment effects on their original scales. The mappings between standardised effects are reported as relative responsiveness ratios.

### *Key findings*

We found evidence that the PHQ-9 was most responsive to change following treatment of the depression scales investigated. For example, a 1-SD unit treatment effect of BDI-II was equivalent to a 1.52-SD unit effect on the PHQ-9 (95% credibility interval 1.17 to 2.05) and a 1.31-SD unit effect on HAMD-17 (95% credibility interval 1.04 to 1.69). This is evidence that the PHQ-9 is more responsive to treatment changes than the BDI, by a factor of 1.52, and the HAMD-17 is more responsive than the BDI, by a factor of 1.31. The finding that the PHQ-9 was superior to the BDI agrees with a previous finding.[55] There was evidence that the generic EuroQol-5 Dimensions (EQ-5D) and SF-36 measures were less sensitive to change than the BDI.

### *Limitations*

Findings from this study are limited by its small size, the unrepresentative sample of trials that were selected and the ability to generalise to other clinical situations.

### *Inter-relation with other parts of the programme*

This study provided evidence that the PHQ-9 was more responsive to change after treatment than other depression measures, supporting its use as an outcome measure in the PANDA RCT. The mapping coefficients were also used in the value-of-information study (aim 1c) as the mapping coefficients can be used to estimate quality-of-life differences when only depression measures have been used in the study.

## Aim 1c: to assess the value of information from carrying out a randomised controlled trial of antidepressants in depression of mild severity

### *Research aims*

Our aims were to develop an economic model that incorporates the severity of depression as part of decision-making processes. This would lead to a recommendation of the most cost-effective threshold above which to prescribe antidepressants and also an estimate of the value of a trial aiming to reduce uncertainty in this decision.

### *Methods*

We extracted trial data from those identified in earlier systematic reviews by Kirsch *et al.*,[20] Fournier *et al.*[22] and Gibbons *et al.*[23] Cipriani *et al.*[56] provided evidence on discontinuation rate in the first 12 weeks of treatment. To address gaps in evidence for our economic model we obtained expert clinical opinion.

### *Analysis*

The model is split into two components. The first was a continuous estimate of HAMD at the end of the initial 12 weeks as a function of the initial score. This was based on a metaregression of the extracted trials reporting treatment effect and baseline depression severity conducted using the Bayesian software WinBUGS version 1.4.3 (MRC Biostatistics Unit, Cambridge, UK).[57] This metaregression estimated proportional treatment and placebo effects on depression severity on the HAMD scale.

This outcome of this model was then categorised into four depression severities: well is 0–7 HAMD, mild is 8–13 HAMD, moderate is 14–18 HAMD, and severe and very severe are > 19 HAMD. These four states formed a Markov model[58] that extrapolated patient progress over a further 2 years in eight 12-week cycles. The HAMD for each state was mapped to the EQ-5D using the results of aim 1b, giving quality-adjusted life-years (QALYs) for each cycle. Total QALYs and costs were calculated and gave the incremental net benefit for each treatment strategy. The expected value of partial perfect information (EVPPI), which is the improvement to decision-making if uncertainty on a selection of input parameters to the model were removed, was used to determine an upper bound on the value of collecting further evidence.[59]

### *Key findings*

The metaregression estimated that patients on antidepressants had an additional 12% (95% credibility interval 3% to 21%) decrease in 6-week HAMD versus placebo. The economic model determined that treating patients with a severity score of $\geq 2$ on HAMD had the highest probability (> 65%) of being cost-effective at a £20,000 willingness-to-pay threshold.

A short-term trial investigating the relation between treatment effect and severity and quality of life in depression patients had an EVPPI of £67.7M over a 10-year time horizon. This suggested that the proposed PANDA trial was potentially cost-effective.

### *Limitations*

There was little evidence on treatment effects in low-severity patients, but our analysis had assumed that the relationship with severity held across the whole range of HAMD scores. We were reliant on clinical opinion for some important values affecting costs. Finally, the EVPPI provides an upper bound on the value of a trial as it assumes the removal of all uncertainty on a subset of parameters. Expected value of sample information estimates the value of reduced uncertainty on a subset and is related to a specified sample size and trial design; expected value of sample information would be necessary for a more accurate assessment of trial value.[60]

### *Inter-relation with other parts of the programme*

We estimated that the PANDA RCT (phase 3 of the programme grant) had the potential to be cost-effective and the absolute expected value of perfect information estimates would vary from approximately £70M to £95M between the models. The metaregression results of previous trials informed our power calculations for the RCT.

# Phase 2: the PANDA cohort study – using both quantitative and qualitative methods

## Aim 2a: estimating a clinically important difference in commonly used self-administered questionnaires for depressive symptoms

### *Research aims*

The aim of this study was to estimate the MCID for the PHQ-9, the BDI-II and the Generalised Anxiety Disorder-7 (GAD-7) using an anchoring method in which participants were asked to retrospectively report improvement or worsening on a GRC question. We also investigated whether or not the MCID varied according to the initial severity of symptoms.

### *Methods for data collection*

The PANDA cohort consisted of patients who had presented to UK primary care surgeries with depressive symptoms or disorder, or depressed mood, during the previous year. Participants had a range of depressive symptoms and were recruited from one population, reducing selection bias. Overall, 7721 patients were sent an information letter in the post and 1470 (19%) replied (*Figure 2*). Of these, 821 were willing to be contacted, 23 (3%) of whom were ineligible. The remaining 798 were contacted to arrange an interview and 563 consented to take part in the cohort study. Data on our measures were collected at four time points. At time 1, 559 people provided data (four could not be contacted), with corresponding figures at follow-up at 2, 4 and 6 weeks of 476 (85%), 443 (79%) and 430 (77%), respectively. For this analysis we used data from 400 participants who gave complete data on all follow-ups.

The participants were asked to rate their own improvement using a GRC question. The GRC was assessed by asking patients '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) and 'I feel a lot worse' (5). These ratings were compared with the changes in the score on the self-administered questionnaires BDI-II, PHQ-9 and GAD-7. This enabled us to calculate the change in scores (on the questionnaires) that corresponded to an improvement in patients' GRC. To assess the reliability of the GRC, the question was completed twice by the participants at each time point. The participants completed the Clinical Interview Schedule – Revised (CIS-R)[61] at baseline only. The data were used to derive the diagnosis of depression and its severity.

### *Analysis*

We analysed data from 400 participants with complete data on the CIS-R, PHQ-9, BDI-II, GAD-7 and GRC. We assessed reliability of the GRC scale by quantifying the two repeated assessments completed by the participant at each follow-up in absolute and relative terms. We used beta regression to estimate the changes in depressive symptoms measured by the BDI-II and the PHQ-9 over three follow-ups in each GRC category and according to three categories of the CIS-R score. This was an improvement on our previous analysis as it allowed us to model variability and means.

### *Key findings*

We estimated the threshold below which the participant was more likely to report feeling better than feeling the same using a ROC analysis. The estimates were provided for three severity bands, determined by the baseline CIS-R score. The average initial scores for the three bands on the PHQ-9 were 4.1, 7.8 and 12.2.

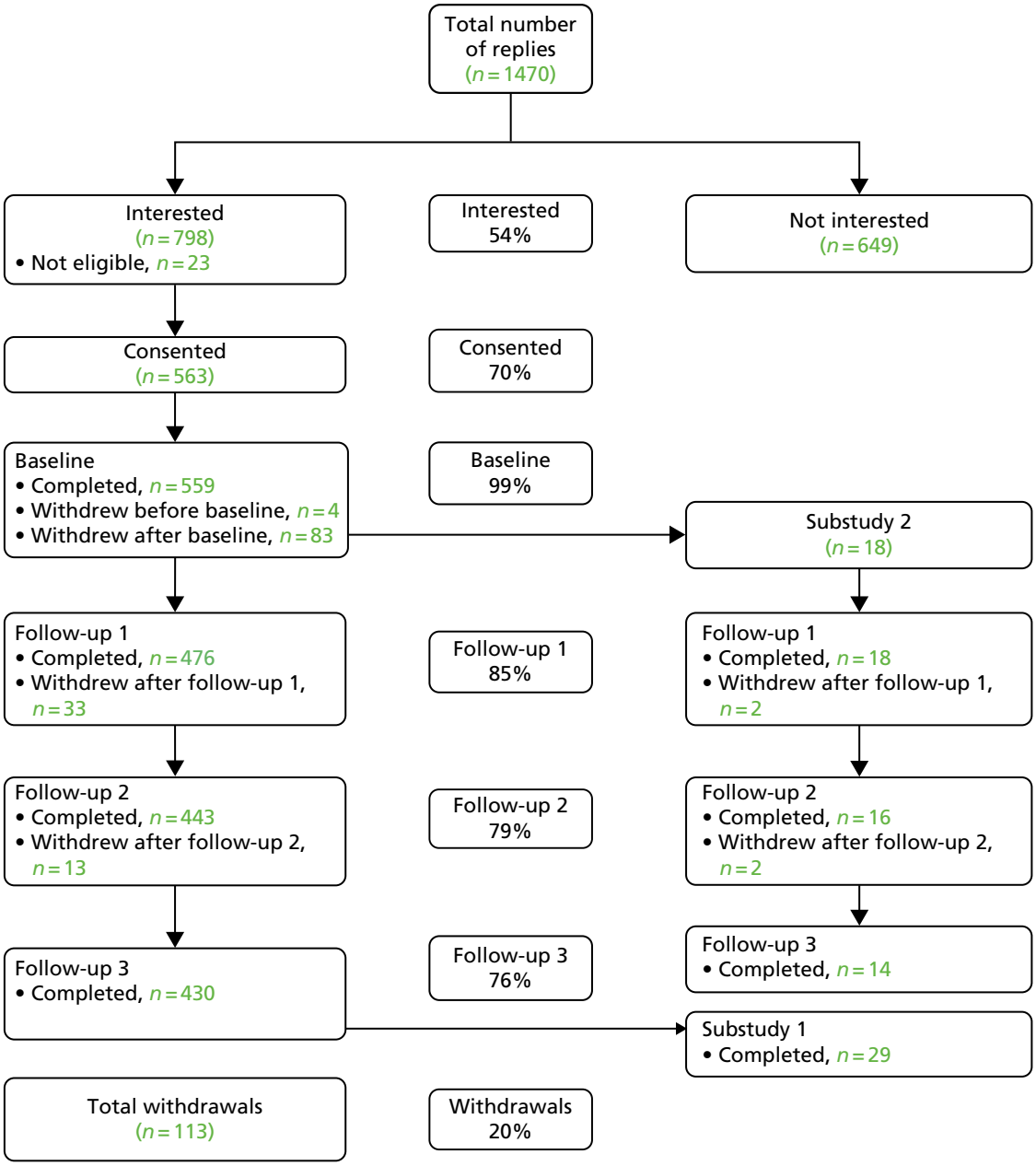



**FIGURE 2** The Consolidated Standards of Reporting Trials (CONSORT) flow diagram: PANDA cohort study.

The range of scores therefore extended our study of MCID to lower ranges of severity than our earlier study (aim 1a, see *Appendix 1*). The MCID as a percentage appeared to increase for the lower severities. For example, in the lowest severity band, the MCID for the PHQ-9 was 48% (95% CI 37% to 65%), whereas in the most severe band it was 19% (95% CI 16% to 24%). There was still considerable uncertainty about the MCID at lower severities. For the GAD-7, the corresponding figures were 72% (95% CI 55% to 97%) and 9% (95% CI 7% to 11%).

### *Limitations*

There was relatively low power in this cohort because there was little change in symptoms between the follow-up points. We had a low response rate in the cohort, but one would not expect any selection bias to affect the estimates. The participants who 'felt the same' on the GRC still had a drop in score and it is not clear why this occurred.

### *Inter-relation with other parts of the programme*

The MCID is essential if we are to give guidance to patients and doctors about whether or not antidepressants should be prescribed. The aim of both MCID studies (aims 1a and 2a) was to develop, for the first time, a patient-centred measure of the change in depressive symptoms required to achieve a clinical benefit. These data will be able to help interpret the results of the PANDA RCT as well as being an output in their own right for other investigators.

Our approach towards estimating the MCID was based on an average within-person change related to improvement. However, the results from RCTs provide an estimate based on a comparison between groups. Applying our MCID estimate to a RCT result therefore rests on a counterfactual argument in which the outcome were that individual to receive a placebo is contrasted with the outcome were that individual to receive the active treatment. In other words, the patient is told 'if you received the treatment you would (on average) be X points lower on the PHQ-9 and (on average) that is a difference that people regard as important'. Using this argument allows clinicians to make treatment recommendations for individual patients under counterfactual arguments resting on the trial's generalisability. We describe the probability that a patient who 'feels better' has a reduction in depression score scale of greater than or equal to the MCID. It is natural to compare the expected benefit from an intervention tested in a RCT with that minimum difference. It gives us an idea of the improvement needed for the patient to perceive any benefit.

## Aim 2b: to investigate the changes reported by patients as they recover from depression

We carried out three studies to investigate this aim. First, we carried out a qualitative investigation of the meaningfulness of the PHQ-9 in determining meaningful symptoms of low mood. We also examined the processing of emotional information and how this varied with depressive symptoms and over time. Emotion-processing is a key abnormality in depression and influenced by antidepressant medication.[14] The second study investigated the variation of emotional face recognition in relation to depressive symptoms. Finally, the third study examined variation in recall of socially rewarding information according to depressive symptoms.

### *Study 1: a qualitative investigation of the meaningfulness of the PHQ-9 in determining meaningful symptoms of low mood*

#### Research aims

To explore differences between the way patients comprehend and map their answer to the options on the questionnaire. A secondary aim was to investigate whether or not patients shift over time in how they comprehend items on the questionnaire or find them problematic to answer, in relation to their own changing symptoms. The substudy also examined the content of responses and their meaning to the participants.

#### Methods of data collection

This was a longitudinal qualitative substudy nested within the PANDA cohort study, which included 18 participants who completed the baseline appointment at the Bristol site. A purposive sampling strategy was used to ensure that there was a range of participants of differing ethnicity and sex and sociodemographic differences were presented. The participants were interviewed using cognitive interviewing techniques at 2, 4 and 6 weeks after their baseline. At each interview the participants were invited to complete the GRC question and the PHQ-9 while thinking aloud what was going through their minds. Non-directive, open verbal probing as well as observation probes were used (e.g. 'You're hesitating; can you tell me why?', which was followed by targeted probes, such as 'What does that term mean to you?'). Forty-eight digitally recorded interviews were recorded and analysed.

### Analysis
The analysis used was consistent to that used in cognitive interview framework analysis.[62] A Microsoft Excel® (Microsoft Corporation, Redmond, WA, USA) grid was created to analyse the digital audio files; the grid contained 18 column headings, each heading denoting 'comprehension' or 'answer mapping' for each item on the PHQ-9. Additional columns summarised the data from the card-sorting exercise and the GRC question. Participants were listed in rows, where each row represented a different time point.

### Key findings
The study provided evidence that the PHQ-9 may be missing the presence and/or intensity of certain symptoms that are meaningful to patients. For instance, participants translated the options on frequency into their own meaningful measure of intensity; for example, 'several days' was used to represent a low level of intensity rather than the actual number of days a certain symptom was present. The triple- or double-barrelled questions were problematic for participants who felt that they could respond differently to each part of the question. For example, item 9 on the PHQ-9 asks if patients have been bothered with 'thoughts that you would be better off dead, or hurting yourself in some way'. The participants regarded the GRC as a good way of summarising their situation overall, in contrast to the PHQ-9, which addressed only some of the important changes.

### Limitations
The cognitive interviewing technique is still developing as a framework and the approaches to analysis of the data collected of cognitive interview data are being debated.[63] This was a small sample looking at a limited range of questions.

### Inter-relation with other parts of the programme
This study helps us to understand some of the limitations of the PHQ-9 from the perspective of patients. It provides evidence that the GRC item has validity from the perspective of the patients and indicates the weaknesses of the PHQ-9 in assessing individual change.

## *Study 2: variation in emotional face recognition and depressive symptom severity*

### Research aims
The aim was to investigate whether or not processing of happy and sad facial expressions was associated with the severity of depressive symptoms, cross-sectionally and longitudinally.

### Methods of data collection
In this study, we examined the data from the computerised facial recognition task that was completed by the PANDA cohort participants ($n = 509$) at baseline, then at 2 and 4 weeks. The participants were presented with 'morphed' faces with varying degrees of emotional intensity. The correct responses were classified as 'hits' and incorrect responses as 'false alarms'. Accuracy and response bias were measures for facial expressions of varying emotional intensities.

### Analysis
Analyses were conducted using multilevel or mixed-effects linear regression models to calculate concurrent and longitudinal associations between hits, false alarms and depressive symptoms separately for happy and sad faces.

### Key findings
For every additional face incorrectly classified as happy (positive emotion bias), concurrent PHQ-9 scores reduced by 0.05 of a point (95% CI –0.10 to 0.002; $p = 0.06$). This association was strongest for more ambiguous facial expressions. There was no evidence for associations between sad face recognition and concurrent depressive symptoms, or between happy or sad face recognition and subsequent depressive symptoms or antidepressant use. We concluded that as the severity of depressive symptoms increased there was a reduced tendency to see positive images but there was no influence on negative images.

### Limitations

The sample excluded people with depression who had not visited their GP and we had a low response rate. However, the inclusion of participants did not depend on emotion recognition, so this is unlikely to have biased any associations between emotion recognition and depressive symptoms. There was little change in our cohort so we cannot exclude the possibility of longitudinal associations between facial expression recognition and depressive symptoms.

### Inter-relation with other parts of the programme

The results indicated that, as depressive symptoms increased, people became less likely to report that an ambiguous facial expression was happy. This has important implications for understanding how people with depression might respond to social circumstances. We demonstrated that this effect occurred over the whole range of depressive symptom severity. Future research could identify whether or not emotion-processing performance could be used to predict response to antidepressants.

## *Study 3: variation in the recall of socially rewarding information and depressive symptom severity*

### Research aims

The aim was to investigate whether depressive symptoms are associated with recall for socially rewarding (positive) or socially critical (negative) information.

### Methods of data collection

This study also used the data from the PANDA cohort and, as in the previous study of facial recognition, positive and negative recall were assessed at three time points: baseline and 2 and 4 weeks. On each occasion, participants were presented with 20 likeable and 20 unlikeable faces on a computer screen in a random order. Participants had to rate whether these were likeable or unlikeable and, after a short gap, were asked to recall any of the words that were presented.

### Analysis

Analyses were conducted using multilevel mixed-effects models to calculate concurrent and longitudinal associations between the number of positive and negative words recalled and depressive symptoms, before and after adjustment for confounders ($n = 524$).

### Key findings

We found evidence for a concurrent association between increased recall of positive words and reduced severity of depressive symptoms: for every increase in two positive words recalled, depressive symptoms reduced by 0.6 (95% CI –1.0 to –0.2) BDI points. There was no evidence of an association between depressive symptoms and negative recall (–0.1, 95% CI –0.5 to 0.3). Longitudinally, we found more evidence that increased positive recall was associated with reduced depressive symptoms than vice versa.

### Limitations

Although the analysis was conducted on the largest sample to date of emotional processing and depressive symptoms, the cohort study had a low response rate, which might have introduced a selection bias. Although different words were used at each time point, after the first assessment participants would have expected the incidental recall task. This could have led to increased recall, but we did not observe this.

## Aim 2c: to investigate disagreement between self-reported improvement and changes in the scores on depressive symptom questionnaires

We used quantitative and qualitative methods to identify those aspects of recovery that are currently missed by questionnaires.

## *Study 1: why are there discrepancies between depressed patients' Global Rating of Change and scores on the PHQ-9 depression module? A qualitative study in primary care*

### Research aims

The aim of this qualitative study was to investigate why there are discrepancies between depressed patients' GRC in their mood and their scores on the PHQ-9. Patients were interviewed regarding the source and meaning of mismatches between their GRC and their PHQ-9 scores.

### Methods of data collection

This study was nested within the larger PANDA cohort study in which participants completed the GRC and a PHQ-9 at four time points, each 2 weeks apart. We examined data from the first 86 participants in Liverpool who had completed all study assessments. 'Mismatch' was defined as a disagreement between a patient's GRC and a meaningful change in their PHQ-9 scores between that time point and the preceding one. We classified a meaningful change as a 15% reduction or increase in scores, based on preliminary MCID estimates from the programme. Of the 86 participants selected, 44 (51%) were identified as cases of mismatch. The 32 participants with the most pronounced mismatch were invited to participate in the qualitative substudy. Qualitative interviews were audiotaped and transcribed with 29 participants. The interview centred on five key topics: experiences of depression, experiences and expectations of treatments, how effective they thought the questionnaires were (e.g. the PHQ-9), reasons for their mismatch and social factors.

### Analysis

Interpretative phenomenological analysis (IPA) was used to guide the analysis. This enabled us to focus on the individual accounts before moving to identify more general themes in the data. All transcripts were coded to identify initial themes, and then further analysed to formulate superordinate and subthemes.

### Key findings

We identified four superordinate themes as explanations for disagreement:

1. There were limitations in the questions asked by the PHQ-9 and a lack of questioning about intensity such that the GRC provided a more accurate assessment of current mental state. The PHQ-9 does not ask about some depressive symptoms, such as interacting with people, lack of libido and inability to cope at work. It also does not enquire about comorbid symptoms such as anxiety, PTSD symptoms and physical illnesses.
2. The impact of recent positive or negative life events could affect their responses but was not captured by the PHQ-9.
3. Variation in mood was 'normal' so was not seen as a global change in mood. Participants had underscored responses in the hope that their symptoms would improve or did not want to admit how they were feeling. Participants sometimes omit items on suicidality to avoid possible intervention.
4. Some participants observed that they found it difficult to recall what they were doing or how they were feeling from one day to the next.

### Limitations

This was a relatively small sample and it is not possible to infer how common the reasons for disagreement might be in a more representative sample. The MCID estimate was based on preliminary results.

### Inter-relation with other parts of the programme

This study helps to further understand some of the limitations of the PHQ-9 in assessing change. It supports the view that the PHQ-9 should not be used alone to assess improvement in individuals. Such self-administered scales need to be supplemented with further clinical assessment. Further clinical assessment is needed if the PHQ-9 is to be used in clinical practice. This study supported the validity of the GRC, but some respondents found the retrospective recall required by the question was difficult.

## *Study 2: changes in self-administered measures of depression severity and patients' own perceptions of changes in their mood – a prospective cohort study*

### Research aims
The aim was to examine the extent to which changes in scores from self-administered depression questionnaires (PHQ-9 and BDI-II) disagree with patients' own self-rated improvement in mood, and investigate factors that influence this relationship.

### Methods of data collection
We used data on the BDI-II and the PHQ-9 and the GRC completed by the PANDA cohort participants at baseline and at the 2-, 4- and 6-week follow-ups.

### Analysis
The change scores for the BDI-II and the PHQ-9 at the 2-, 4- and 6-week follow-ups were calculated by subtraction from the previous time point. We used a MCID of 20% to create categories of meaningful improvement, no change and deterioration that could be compared with the GRC. We used logistic regression models to test whether or not anxiety symptoms, mental and physical health-related quality of life, negative life events, and social support influenced response to the GRC after adjustment for the change scores on the BDI-II or the PHQ-9.

### Key findings
About half of the patients exhibited disagreement between their response on the GRC and the categories of meaningful change that we calculated. For the PHQ-9 we found that 51% (95% CI 46% to 55%) showed disagreement and for the BDI-II we found that 55% (95% CI 51% to 60%) showed disagreement. We also found that patients with more severe anxiety symptoms were less likely to report feeling better on the GRC, having taking account of the change in depressive symptoms. Patients with a better mental health-related quality of life were more likely to report feeling better in a similar analysis. Thus, anxiety and health-related quality of life contribute to the perception of improvement over and above any change in depressive symptoms.

### Limitations
The PANDA cohort had a low response rate that might have introduced a selection bias. However, as our selection of patients did not depend on any of the exposure variables, it is unlikely to have biased the associations we have reported.

### Inter-relation with other parts of the programme
The results of this quantitative study supported our finding from the qualitative PANDA studies that clinicians working in primary care and other clinical settings should be cautious in interpreting changes in questionnaire scores without further clinical assessment. The study indicated areas where depressive symptoms questionnaires are not assessing aspects of mental health important to patients. In a RCT, any variation between individuals should not affect the comparison as the randomisation should lead to comparable groups.

# Phase 3: the PANDA randomised controlled trial

## Research aims

The aim of the RCT was to investigate the severity and duration of depressive symptoms that are associated with a clinically important response to sertraline and cost-effectiveness (compared with placebo) in people with depression who present to primary care. The main hypothesis was that response to antidepressants would increase with both severity and duration of depressive symptoms. Our primary analysis was a comparison between sertraline and placebo at 6 weeks.

## Methods for data collection

We used broad and pragmatic inclusion criteria, recruiting people who had sought treatment for depressive symptoms of any severity or duration in primary care. The key entry criterion was that GPs and/or patients were uncertain about the potential benefits of an antidepressant and we did not set any severity or duration thresholds as exclusions. Patients were recruited from primary care surgeries in four UK sites (i.e. Bristol, London, Liverpool and York) and identified by GPs, who either invited patients during a consultation or conducted a database search and then sent an invitation in the post. Participants were randomised to 100 mg of sertraline or an identical placebo and followed up at 2, 6 and 12 weeks.

*Figure 3* is a flow diagram of the progress through the trial.

## Analysis

The primary outcome was the PHQ-9 at the 6-week follow-up. Interactions terms of a realistic size, that are smaller than the main effect, require very large sample sizes for adequate power. As a result, we modelled the treatment effect on log-transformed PHQ-9 scores (continuous outcome) using an intention-to-treat analysis. The exponentiated regression coefficient is the proportional (or percentage) change in PHQ-9 scores between randomised groups. Evidence of a treatment effect using a proportional model implies that the treatment effect expressed as a mean difference would increase with severity. In sensitivity analyses we fitted an additive model using absolute depression scores (non-logged PHQ-9) and calculated an interaction between treatment allocation and baseline CIS-R depression severity score. However, we expected the power of this analysis to be low.

Secondary outcomes at 2, 6 and 12 weeks were depressive symptoms and remission assessed using the PHQ-9 and the BDI-II, generalised anxiety disorder symptoms, mental and physical health-related quality of life and self-reported global improvement. We used linear multilevel models for repeated measures of continuous secondary outcomes at 2, 6 and 12 weeks (PHQ-9, BDI-II, GAD-7 and Short Form questionnaire-12 items physical and mental health-related quality of life). Logistic multilevel models were calculated for repeated measures of binary secondary outcomes at 2, 6, and 12 weeks (remission on the PHQ-9, the BDI-II and feeling better on the GRC scale).

We undertook a cost-effectiveness analysis from the perspective of the NHS and Personal and Social Services alongside the PANDA RCT. Quality-of-life data were collected at baseline and 2, 6 and 12 weeks post randomisation using EQ-5D-5L, from which we calculated QALYs. Costs were collected using patient records and from resource use questionnaires administered at each follow-up interval. Differences in mean costs and mean QALYs and net monetary benefits were estimated. Our primary analysis used net monetary benefit regressions to identify any interaction between the cost-effectiveness of sertraline and subgroups

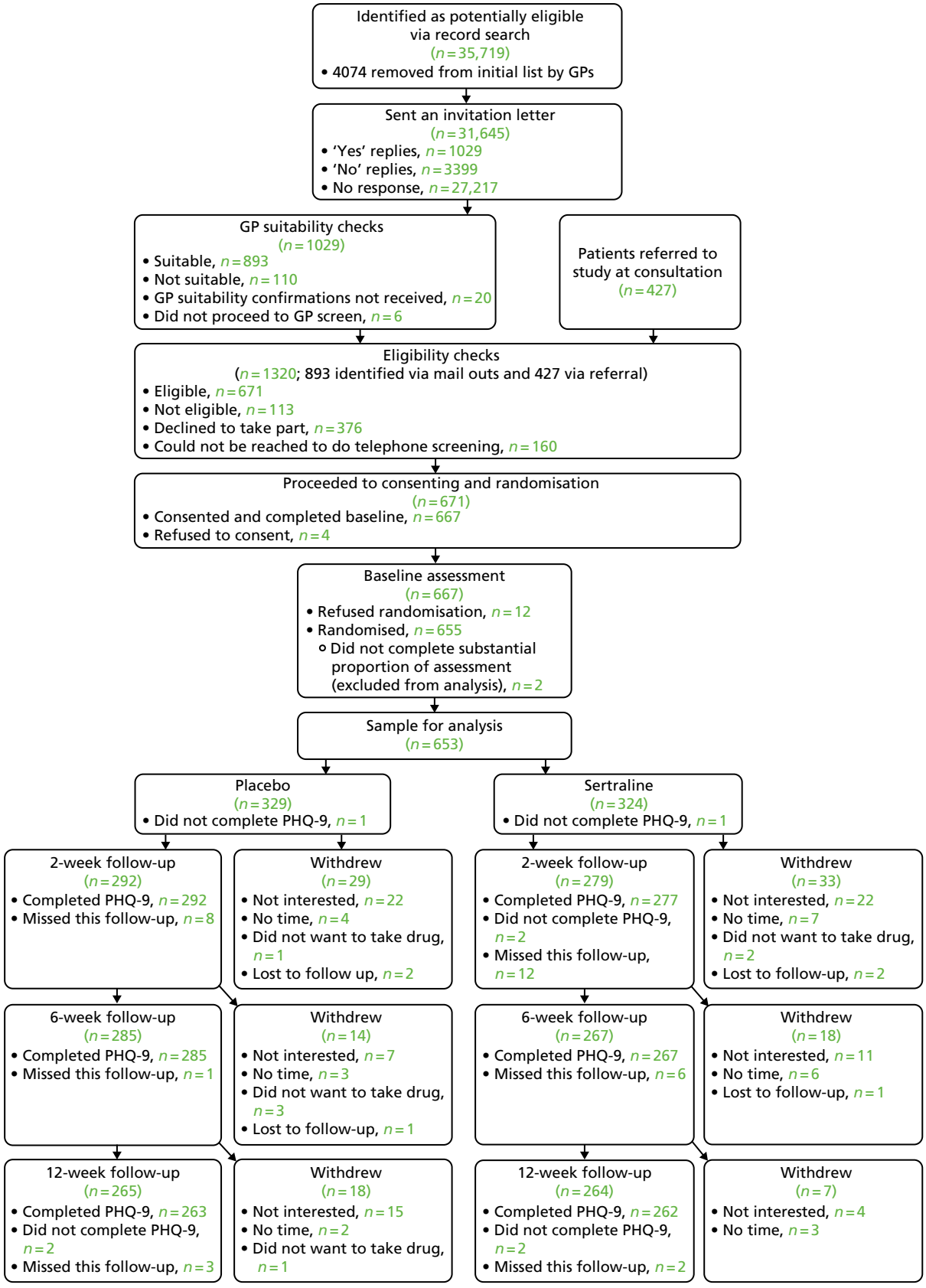



**FIGURE 3** The Consolidated Standards of Reporting Trials (CONSORT) flow diagram: PANDA RCT.

defined by baseline symptom severity (0 to 11; 12 to 19; ≥ 20 on the CIS-R) and, separately, duration of symptoms (greater or less than 2 years' duration). A secondary analysis estimated the cost-effectiveness of sertraline versus placebo. In sensitivity analyses, we (1) performed a complete-case analysis to check the robustness of our findings to missing data, (2) examined the impact of excluding costs that were not judged to be directly related to the treatment of depression and (3) excluded all secondary care costs from total NHS and Personal and Social Services costs to assess whether or not our findings were robust to infrequent but expensive hospitalisations.

## Key findings

We found no evidence that the antidepressant sertraline reduced depressive symptoms at 6 weeks. In the sertraline group, PHQ-9 scores were 5% (95% CI –7% to 15%; $p = 0.41$) lower than those in the placebo group. In the sensitivity analyses using additive models, there was no evidence of an interaction with severity or duration of depressive symptoms with treatment effect, but these analyses would have lacked statistical power.

Of the secondary outcomes, there was strong evidence that sertraline reduced anxiety symptoms (GAD-7 score reduced by 17%, 95% CI 9% to 25%; $p < 0.000046$) and improved mental but not physical health-related quality of life as well as self-reported global improvement. There was weak evidence that depressive symptoms were reduced by sertraline at 12 weeks for both the PHQ-9 and the BDI-II. Given our findings, we also investigated whether or not the treatment effect for anxiety symptoms was influenced by baseline severity. We found no evidence that the effect of sertraline on anxiety symptoms varied according to the severity of anxiety or depressive symptoms. The number needed to treat in order to feel better according to our self-reported global improvement question was 8.5 (95% CI 5.2 to 22.1) people at 6 weeks and 6.4 (95% CI 4.6 to 10.3) people at 12 weeks.

There was no evidence of an association between the baseline severity of depressive symptoms and the cost-effectiveness of sertraline. Compared with patients with low symptom severity, the expected net benefits in patients with moderate symptoms were £64 (95% CI –£312 to £441) and the expected net benefits in patients with high symptom severity were –£51 (95% CI –£389 to £287). Patients who had a longer history of depressive symptoms at baseline had lower expected net benefits from sertraline than those with a shorter history; however, the difference was uncertain (–£132, 95% CI –£431 to £167). In the secondary analysis, patients treated with sertraline had higher expected net benefits (£118.37, 95% CI –£23.39 to £260.14) than those in the placebo group. Sertraline had a high probability (> 90%) of being cost-effective if the health system was willing to pay at least £20,000 per QALY gained.

## Limitations

We had broad inclusion criteria and some participants had very few symptoms. This may have reduced the treatment effect, but our methods of analysis using a proportional approach should have helped to take account of this. There was attrition of nearly 20% by 12 weeks, although this did not differ by study arm, and when we investigated the impact of missing data, this did not appear to explain the findings. We had limited statistical power to explore interactions between treatment response and symptom severity or duration. It is possible that subgroups in which sertraline was more cost-effective might have become more evident with a larger sample size or longer follow-up.

## Inter-relation with other parts of the programme

The RCT did not find evidence of an early antidepressant effect of sertraline on depressive symptoms in the population studied. There was, however, evidence that sertraline reduced anxiety symptoms and was

more likely to lead to a clinically important benefit. The results of the trial and those from the MCID can be used to provide some initial guidance about the likely cost-effectiveness of sertraline and other SSRI antidepressants in primary care.

## Conclusion

We chose the PHQ-9 as our primary outcome in the PANDA trial. The PHQ-9 is widely used in primary care and there is evidence that it is better at detecting changes in depressive symptoms after treatment than other measures.[37,55] It also avoids the observer bias that affects clinician-rated HAMD and MADRS scores. However, our qualitative research identified a number of reasons for disagreement between the PHQ-9 and the self-reported GRC, as well as indicating that up to 50% of patients might show disagreement between self-reported change and the results of questionnaires. Our findings indicate that the processes and motivations behind completing the PHQ-9 are complex and influenced by ongoing physical, social and emotional issues. Our findings suggest that PHQ-9 and, by implication, other self-administered questionnaires should not be used alone to assess improvement or deterioration. Their use should be supplemented with further clinical assessment and the use of more open-ended questions.

The MCID is the smallest change in symptoms that is considered clinically worthwhile by the patient. Our MCID research in phases 1 and 2 has enabled us to develop appropriate analytical methods for estimating the MCID and to provide values for the MCID in a primary care population for the first time. We can apply our MCID estimates to the results of the PANDA trial, but we acknowledge that our estimates of MCID are still uncertain.

When we initially formulated our research questions we assumed that the treatment effect varied according to depression severity but that the MCID was a fixed value, irrespective of depression severity. Our results, if anything, now point in the opposite direction, at least when we estimate treatment effects and MCID using a proportional approach. Our results suggest that sertraline has a similar (proportional) effect size over the whole range of depression (and anxiety) severity and it is the MCID that changes and gets larger, proportionally, at lower levels of severity. We have found that, at higher levels of severity, the proportional approach works better for the MCID. The proportional approach also has attractions when analysing clinical trial data. It avoids the assumption that the same absolute treatment effect is observed regardless of whether a person scores 5 or 25 on a scale. This seems unlikely and a proportional reduction approach appears more plausible as well as providing a better statistical fit to the data.

Contrary to our initial hypothesis in the PANDA trial, we found no evidence of a clinically important effect of sertraline on depressive symptoms. We found a 5% reduction in the sertraline group at 6 weeks, and this is considerably smaller than the MCID estimates for the PHQ-9 we have obtained. We cannot exclude the possibility that sertraline led to a clinically important improvement at 12 weeks, as we found a 13% (95% CI 3% to 21%) reduction in PHQ-9, but this is still well below our MCID. In contrast we found strong evidence that sertraline reduced anxiety symptoms on the GAD-7, with a reduction of 21% (95% CI 11% to 30%) at 6 weeks. This is consistent with some of our estimates of the MCID for GAD-7 (see *Appendix 4*) and suggests that this change is clinically important. We found insufficient evidence of an interaction between the cost-effectiveness of sertraline and severity or symptom duration that GPs could use to efficiently target prescribing. There was no evidence of a substantial treatment effect of sertraline on quality of life, as measured by the EQ-5D-5L. However, sertraline is an inexpensive intervention that has a high probability of being cost-effective compared with placebo across primary care patients with depression or low mood.

Our MCID estimates are based on an average within-person change related to improvement; however, a RCT compares groups. Application of our results has therefore required a counterfactual argument in which researchers compare the same individual(s) who receive placebo but who might have received the active treatment. The MCID estimated from a within-person calculation can then be applied to the between-group differences in a clinical trial.

Our estimates of MCID could be used to guide decisions about whether or not a treatment will benefit an individual. For this, one needs to be able to predict the likely score for that person on the proposed outcome measure were they not to receive that treatment. In other words, we need to know the likely value of an individual's PHQ-9 or GAD-7 score at follow-up, 6 or 12 weeks later, if they were to receive a placebo. The expected value of the placebo at follow-up can then be used to determine the appropriate initial value for the MCID and thus decide if the proportional reduction expected from a treatment would be larger than the MCID for such a person. For this to be feasible, we would need more precise estimates of MCID and also the ability to predict future scores of patients on the basis of their clinical and other characteristics.

Finally, we can make some very approximate estimate of what proportion of participants in the PANDA cohort would probably have benefited from treatment. From our results in *Appendix 4*, it is clear that those with a GAD-7 score of 3 at 6 weeks have a MCID of about 50%. It is highly unlikely that those individuals would have experienced any benefit from sertraline. In the PANDA RCT, about 30% of participants scored $\leq 3$ at 6 weeks. However, we cannot conclude this with any confidence at this stage. We do not know the distribution of symptoms in those receiving antidepressants in the UK and our estimates of MCID are approximate. Our overall results are reassuring in indicating that, on average, patients in the PANDA RCT are benefiting from sertraline. However, it is probable that a substantial proportion or patients receiving antidepressants are not experiencing any individual benefit. For clinicians to be confident about recommending treatment to patients, we need accurate information on individualised treatment effects and the outcome without treatment as well as MCID. Of course, any recommendations for treatment will also have to take account of any risks and adverse effects that result from the treatment as well as patient preference.

## Recommendations for future research

Our finding that sertraline seems to be effective for anxiety but not depressive symptoms has a potential implication for understanding the mechanisms of antidepressant treatment as well as the clinical benefit that patient's will experience.

### *Research recommendation 1*

Future research into the mechanism of action of antidepressants should examine the biology of anxiety symptoms.

The result of the RCT also questions the reliance of current clinical guidelines on existing placebo-controlled studies that have been conducted largely for regulatory purposes. Cipriani *et al.*'s review[28] highlights the poor quality of the existing research. Antidepressants are commonly used medications and it is concerning that we still have a number of outstanding questions about their efficacy and clinical indication many years after they were introduced. The use of behavioural tasks such as face recognition and memory for words might be a useful way to investigate these mechanistic aspects.

### *Research recommendation 2*

Future studies should investigate the clinical effectiveness of antidepressants for anxiety disorders in UK primary care population.

We would recommend that future investigation of antidepressant efficacy should have longer follow-ups to see if there are longer-term benefits for depressive symptoms as well as anxiety symptoms. We would encourage use of more detailed outcome measures, using self-reported information, to ensure that the whole range of symptoms that are common in depression and anxiety are studied. The use of self-reported improvement (GRC) seems a valuable outcome measure in clinical trials.

### *Research recommendation 3*

Further investigation of minimal clinically important differences.

Further research is needed to investigate the size of the MCID and the factors that might influence whether or not patients report improvement. We need more precise estimates to guide decision-making. We have provided evidence that patients' reporting of feeling better can be affected by various other factors, such as anxiety and physical changes. Further investigation of this will also help inform how MCIDs could be used clinically to provide treatment recommendations.

## Implications for practice and any lessons learned

The PHQ-9 and similar self-administered questionnaires should not be used alone in assessing improvement or deterioration. It is important to supplement such standardised measures with a clinical assessment.

Sertraline is effective and cost-effective in reducing anxiety symptoms such as worry and restlessness in the first 6 weeks of treatment in people who present with depressive symptoms. Any effect on depressive symptoms takes longer to emerge; although an improvement in anxious symptoms in someone presenting with depressive symptoms could lead to a clinical benefit. Patients who present to primary care with depressive symptoms have a wide range of severity of symptoms. Overall, this population is likely to benefit from SSRI antidepressants. Our findings support the prescription of SSRI antidepressants in a wider group of participants than previously thought, including those who do not meet diagnostic criteria for depression or generalised anxiety disorder, especially when anxiety symptoms such as worry and restlessness are present.

## Patient and public involvement

Paul Lanham, a service user and a co-applicant, was also a member of the independent steering committee during phase 3 of the programme and was involved in PANDA for over 6 years. Paul Lanham and Derek Riozzie have also contributed to PANDA annual meetings where all co-applicants and researchers discuss progress, review the protocols and discuss any findings. They made important contributions to the discussion and influenced the interpretation of the results and decisions about study design. All study documentations have been revised and commented on by Paul, Derek and the user group co-ordinated by Derek at Liverpool University. In addition, we enlisted the support of the North London Service Users Research Forum (SURF). The SURF was co-founded in 2007 by service users and clinical academic psychiatrists at University College London to provide meaningful consultation on research. It has 12 members with mental health problems. Since 2007, it has consulted on > 50 projects and SURF members have been invited to join steering/management groups on many of these. As a result, the group is very experienced and confident about the advice and input they provide; their comments on the trial paperwork have been invaluable. The letter templates, patient information sheets and the questionnaire were amended to reflect the patient and public involvement (PPI) feedback. We also consulted on the protocol concerning self-harm or risk of self-harm, which we used if patients reported this in the course of the cohort and RCT. Having close involvement of the PPI for the duration of the programme (i.e. over 6 years) has been invaluable for its success. It has also enabled us to build on it and we have recruited a PANDA RCT participant to represent PPI on a different depression trial: ANTidepressants to prevent reLapse in dEpRession (ANTLER).[64] We plan to carry on using the services users' comments in the design, documentation and analysis of any future studies.

# Acknowledgements

We are grateful to all the patients who took part in the PANDA studies. We would like to thank the GPs and GP surgery staff who supported recruitment for this research. We have been supported by the following Clinical Research Networks (CRNs): North Thames CRN, CRN North West London, CRN South London, North West Coast CRN, Greater Manchester CRN, West Midlands CRN, West of England CRN, Yorkshire and Humber CRN and North East and Cumbria CRN.

We would like to acknowledge the particular input of the CRN research nurses and CSOs: Dawn Adams, Heather Tinker, Lynsey Wilson, Tara Harvey, Khatiba Raja, Zara Prem, Beena Bauluck, Yvonne Foreshaw, Cynthia Sajani, Jahnese Maya, Anna Townsend-Rose, Emily Clare, Rachel Nixon, Pam Clark and Irene Sambath.

We would also like to acknowledge Vivien Jones for providing administrative support at the Bristol site; Rebecca Rawlinson at the Liverpool site; Bryony Thomson and Yvonne Donkor at University College London; and Wendy Cattle at York. Jodi Prem was instrumental in recruitment at York.

We also thank Carolyn Chew-Graham, Ian Anderson, Anne Rogers, Evan Kontopantelis, Paul Lanham, Christopher Williams, Richard Bying and Obi Ukoumunne for generously agreeing to sit on the Trial Steering Committee and Data Monitoring Committee.

## Ethics approval and sponsorships

For the PANDA cohort study, ethics approval was obtained from National Research Ethics Service (NRES) Committee South West-Central Bristol. The University of Bristol acted as sponsor for the study.

The PANDA RCT was approved by NRES Committee East of England – Cambridge South (reference number 13/EE/0418). The Joint Research Office, University College London acted as sponsor for the RCT.

## Contributions of authors

**Larisa Duffy** was the programme manager and with **Gemma Lewis** was responsible for drafting the report.

**Larisa Duffy, Gemma Lewis, Anthony Ades, Ricardo Araya, Jessica Bone, Sally Brabyn, Katherine Button, Rachel Churchill, Tim Croudace, Catherine Derrick, Padraig Dixon, Christopher Dowrick, Christopher Fawsitt, Louise Fusco, Simon Gilbody, Catherine Harmer, Catherine Hobbs, William Hollingworth, Vivien Jones, Tony Kendrick, David Kessler, Naila Khan, Daphne Kounali, Paul Lanham, Alice Malpass, Marcus Munafo, Jodi Pervin, Tim Peters, Derek Riozzie, Jude Robinson, George Salaminios, Debbie Sharp, Howard Thom, Laura Thomas, Nicky Welton, Nicola Wiles, Rebecca Woodhouse** and **Glyn Lewis** contributed to the constituent papers in the appendices. The papers included as appendices each have its own lists of authors.

**Anthony Ades, Ricardo Araya, Rachel Churchill, Tim Croudace, Christopher Dowrick, Simon Gilbody, William Hollingworth, Tony Kendrick, David Kessler, Paul Lanham, Alice Malpass, Tim Peters, Jude Robinson, Debbie Sharp, Nicky Welton, Nicola Wiles** and **Glyn Lewis** were responsible for the original proposal and for securing funding.

**Glyn Lewis** was chief investigator of the programme and had clinical responsibly for the RCT recruitment at the London site.

All authors have provided substantial contributions to the conception and design of the PANDA programme and interpretation of data and had input into drafting the report and/or revising it critically for important intellectual content. All authors have given final approval of the version to be published.

## Data-sharing statement

All data requests should be submitted to the corresponding author for consideration. Access to available anonymised data may be granted following review.

## Patient data

This work uses data provided by patients and collected by the NHS as part of their care and support. Using patient data is vital to improve health and care for everyone. There is huge potential to make better use of information from people's patient records, to understand more about disease, develop new treatments, monitor safety, and plan NHS services. Patient data should be kept safe and secure, to protect everyone's privacy, and it's important that there are safeguards to make sure that it is stored and used responsibly. Everyone should be able to find out about how patient data are used. #datasaveslives You can find out more about the background to this citation here: https://understandingpatientdata.org.uk/data-citation.

# References


1. World Health Organization. *Depression: Key Facts*. 2018. URL: www.who.int/news-room/fact-sheets/detail/depression (accessed 15 August 2019).
2. NHS Digital. *Prescription Cost Analysis – England, 2018 [PAS]*. URL: https://digital.nhs.uk/data-and-information/publications/statistical/prescription-cost-analysis/2018 (accessed 15 August 2019).
3. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry* 1960;**23**:56–62. https://doi.org/10.1136/jnnp.23.1.56
4. National Institute for Health and Care Excellence. *Depression: Management of Depression in Primary and Secondary Care. Clinical Guideline 23*. London: NICE; 2004.
5. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol* 1991;**59**:12–19. https://doi.org/10.1037/0022-006X.59.1.12
6. McMillan D, Gilbody S, Richards D. Defining successful treatment outcome in depression using the PHQ-9: a comparison of methods. *J Affect Disord* 2010;**127**:122–9. https://doi.org/10.1016/j.jad.2010.04.030
7. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* 1989;**10**:407–15. https://doi.org/10.1016/0197-2456(89)90005-6
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* 2007;**22**:1596–602. https://doi.org/10.1007/s11606-007-0333-y
9. National Collaborating Centre for Mental Health. *Depression: The Treatment and Management of Depression in Adults (Updated Edition)*. Leicester and London: The British Psychological Society and The Royal College of Psychiatrists; 2010.
10. Löwe B, Unützer J, Callahan CM, Perkins AJ, Kroenke K. Monitoring depression treatment outcomes with the patient health questionnaire-9. *Med Care* 2004;**42**:1194–201. https://doi.org/10.1097/00005650-200412000-00006
11. Jacobson N, Greenley D. What is recovery? A conceptual model and explication. *Psychiatr Serv* 2001;**52**:482–5. https://doi.org/10.1176/appi.ps.52.4.482
12. Ridge D, Ziebland S. 'The old me could never have done that': how people give meaning to recovery following depression. *Qual Heal Res* 2006;**16**:1038–53. https://doi.org/10.1177/1049732306292132
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* 2010;**60**:e231–8. https://doi.org/10.3399/bjgp10X502119
14. Harmer CJ, Goodwin GM, Cowen PJ. Why do antidepressants take so long to work? A cognitive neuropsychological model of antidepressant drug action. *Br J Psychiatry* 2009;**195**:102–8. https://doi.org/10.1192/bjp.bp.108.051193
15. Beck TA, Ruch J, Shaw BF, Emery G. *Cognitive Therapy of Depression*. New York, NY: Guilford Press; 1987.
16. Beck AT, Steer RA, Brown GK. *Manual for the Beck Depression Inventory-II*. San Antonio, TX: Psychological Corporation; 1996.

17. Zigmond A, Snaith R. The hospital anxiety and depression scale. *Acta Psychiatr Scand* 1983;**67**:361–70. https://doi.org/10.1111/j.1600-0447.1983.tb09716.x

18. Cameron IM, Crawford JR, Lawton K, Reid IC. Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care. *Br J Gen Pract* 2008;**58**:32–6. https://doi.org/10.3399/bjgp08X263794

19. National Institute for Health and Care Excellence. *Management of Depression in Primary and Secondary Care. Clinical Guidelines 23*. London: NICE; 2010.

20. Kirsch I, Deacon BJ, Huedo-Medina TB, Scoboria A, Moore TJ, Johnson BT. Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. *PLOS Med* 2008;**5**:e45. https://doi.org/10.1371/journal.pmed.0050045

21. Khan A, Leventhal RM, Khan SR, Brown WA. Severity of depression and response to antidepressants and placebo: an analysis of the Food and Drug Administration database. *J Clin Psychopharmacol* 2002;**22**:40–5. https://doi.org/10.1097/00004714-200202000-00007

22. Fournier JC, DeRubeis RJ, Hollon SD, Dimidjian S, Amsterdam JD, Shelton RC, Fawcett J. Antidepressant drug effects and depression severity: a patient-level meta-analysis. *JAMA* 2010;**303**:47–53. https://doi.org/10.1001/jama.2009.1943

23. Gibbons RD, Hur K, Hendricks Brown C, Davis JM, Mann JJ. 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* 2012;**69**:572–9. https://doi.org/10.1001/archgenpsychiatry.2011.2044

24. Rabinowitz J, Werbeloff N, Mandel FS, Menard F, Marangell L, Kapur S. Initial depression severity and response to antidepressants v. placebo: patient-level data analysis from 34 randomised controlled trials. *Br J Psychiatry* 2016;**209**:427–8. https://doi.org/10.1192/bjp.bp.115.173906

25. Furukawa TA, Maruo K, Noma H, Tanaka S, Imai H, Shinohara K, *et al.* Initial severity of major depression and efficacy of new generation antidepressants: individual participant data meta-analysis. *Acta Psychiatr Scand* 2018;**137**:450–8. https://doi.org/10.1111/acps.12886

26. Barbui C, Cipriani A, Patel V, Ayuso-Mateos JL, van Ommeren M. Efficacy of antidepressants and benzodiazepines in minor depression: systematic review and meta-analysis. *Br J Psychiatry* 2011;**198**:11–16, sup 1. https://doi.org/10.1192/bjp.bp.109.076448

27. Zimmerman M, Posternak MA, Chelminski I. Symptom severity and exclusion from antidepressant efficacy trials. *J Clin Psychopharmacol* 2002;**22**:610–14. https://doi.org/10.1097/00004714-200212000-00011

28. Cipriani A, Furukawa TA, Salanti G, Chaimani A, Atkinson LZ, Ogawa Y, *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* 2018;**391**:1357–66. https://doi.org/10.1016/S0140-6736(17)32802-7

29. World Health Organization. *Classification of Mental and Behavioural Disorders*. Geneva: World Health Organization; 1992.

30. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*: DSM-IV. 4th edn. Washington, D.C.: American Psychiatric Association Publishing; 1994.

31. Rai D, Skapinakis P, Wiles N, Lewis G, Araya R. Common mental disorders, subthreshold symptoms and disability: longitudinal study. *Br J Psychiatry* 2010;**197**:411–12. https://doi.org/10.1192/bjp.bp.110.079244

32. Broadhead WE, Blazer DG, George LK, Tse CK. Depression, disability days, and days lost from work in a prospective epidemiologic survey. *JAMA* 1990;**264**:2524–8. https://doi.org/10.1001/jama.1990.03450190056028

33. de Lima MS, Hotoph M, Wessely S. The efficacy of drug treatments for dysthymia: a systematic review and meta-analysis. *Psychol Med* 1999;**29**:1273–89. https://doi.org/10.1017/S0033291799001324

34. de Lima MS, Moncrieffe JA. Review: antidepressant drugs are effective in dysthymia. *Evid Based Ment Health* 1998;**1**:111. https://doi.org/10.1136/ebmh.1.4.111

35. Anderson IM, Nutt DJ, Deakin JF. Evidence-based guidelines for treating depressive disorders with antidepressants: a revision of the 1993 British Association for Psychopharmacology guidelines. *J Psychopharmacol* 2000;**14**:3–20. https://doi.org/10.1177/026988110001400101

36. Button KS, Kounali D, Thomas L, Wiles NJ, Peters TJ, Welton NJ, *et al.* Minimal clinically important difference on the Beck Depression Inventory–II according to the patient's perspective. *Psychol Med England* 2015;**45**:3269–79. https://doi.org/10.1017/S0033291715001270

37. Kounali DZ, Button KS, Lewis G, Ades AE. The relative responsiveness of test instruments can be estimated using a meta-analytic approach: an illustration with treatments for depression. *J Clin Epidemiol* 2016;**77**:68–77. https://doi.org/10.1016/j.jclinepi.2016.03.005

38. Thom H, Jackson C, Welton N, Sharples L. Using parameter constraints to choose state structures in cost-effectiveness modelling. *PharmacoEconomics* 2017;**35**:951–62. https://doi.org/10.1007/s40273-017-0501-9

39. Malpass A, Dowrick C, Gilbody S, Robinson J, Wiles N, Duffy L, Lewis G. Usefulness of PHQ-9 in primary care to determine meaningful symptoms of low mood: a qualitative study. *Br J Gen Pract* 2016;**66**:e78–84. https://doi.org/10.3399/bjgp16X683473

40. Bone JK, Lewis G, Button KS, Duffy L, Harmer CJ, Munafò MR, *et al.* Variation in recognition of happy and sad facial expressions and self-reported depressive symptom severity: a prospective cohort study. *J Affect Disord* 2019;**257**:461–9. https://doi.org/10.1016/j.jad.2019.06.025

41. Lewis G, Kounali DZ, Button KS, Duffy L, Wiles NJ, Munafò MR, *et al.* Variation in the recall of socially rewarding information and depressive symptom severity: a prospective cohort study. *Acta Psychiatr Scand* 2017;**135**:489–98. https://doi.org/10.1111/acps.12729

42. Robinson J, Khan N, Fusco L, Malpass A, Lewis G, Dowrick C. 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* 2017;**7**:e014519. https://doi.org/10.1136/bmjopen-2016-014519

43. Salaminios G, Duffy L, Ades A, Araya R, Button KS, Churchill R, *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 a randomised controlled trial. *Trials* 2017;**18**:496. https://doi.org/10.1186/s13063-017-2253-4

44. Lewis G, Duffy L, Ades A, Amos R, Araya R, Brabyn S, *et al.* The clinical effectiveness of sertraline in primary care and the role of depression severity and duration (PANDA): a pragmatic, double-blind, placebo-controlled randomised trial [published online ahead of print September 19 2019]. *Lancet* 2019. https://doi.org/10.1016/S2215-0366(19)30366-9

45. Hollingworth W, Fawsitt CG, Dixon P, Duffy L, Araya R, Peters TJ, *et al.* Cost-effectiveness of sertraline in primary care according to initial severity and duration of depressive symptoms: findings from the PANDA RCT [published online ahead of print November 27 2019]. *PharmacoEconomics Open* 2019. https://doi.org/10.1007/s41669-019-00188-5

46. Lewis GH. *PANDA Analysis Plan*. London: Division of Psychiatry, University College London; 2017. URL: http://discovery.ucl.ac.uk/10041458/ (accessed 19 September 2019).

47. Thomas L, Mulligan J, Mason V, Tallon D, Wiles N, Cowen P, *et al.* GENetic and clinical predictors of treatment response in depression: the GenPod randomised trial protocol. *Trials* 2008;**9**:29. https://doi.org/10.1186/1745-6215-9-29

48. Thomas LJ, Abel A, Ridgway N, Peters T, Kessler D, Hollinghurst S, *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;**33**:312–19. https://doi.org/10.1016/j.cct.2011.10.016

49. Baxter H, Winder R, Chalder M, Wright C, Sherlock S, Haase A, *et al.* Physical activity as a treatment for depression: the TREAD randomised trial protocol. *Trials* 2010;**11**:105. https://doi.org/10.1186/1745-6215-11-105

50. Hamilton M. Development of a rating scale for primary depressive illness. *Br J Soc Clin Psychol* 1967;**6**:278–96. https://doi.org/10.1111/j.2044-8260.1967.tb00530.x

51. Montgomery SA, Åsberg M. A new depression scale designed to be sensitive to change. *Br J Psychiatry* 1979;**134**:382–9. https://doi.org/10.1192/bjp.134.4.382

52. Brooks R. EuroQol: the current state of play. *Health Policy* 1996;**37**:53–72. https://doi.org/10.1016/0168-8510(96)00822-6

53. Stewart AD, Hays RD, Ware JE. The MOS short-form General Health Survey. *Med Care* 1988;**26**:724–32. https://doi.org/10.1097/00005650-198807000-00007

54. Lu G, Kounali D, Ades AE. Simultaneous multioutcome synthesis and mapping of treatment effects to a common scale. *Value Health* 2014;**17**:280–7. https://doi.org/10.1016/j.jval.2013.12.006

55. Titov N, Dear BF, McMillan D, Anderson T, Zou J, Sunderland M. Psychometric comparison of the PHQ-9 and BDI-II for measuring response during treatment of depression. *Cogn Behav Ther* 2011;**40**:126–36. https://doi.org/10.1080/16506073.2010.550059

56. Cipriani A, Furukawa TA, Salanti G, Geddes JR, Higgins JP, Churchill R, *et al.* Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet* 2009;**373**:746–58. https://doi.org/10.1016/S0140-6736(09)60046-5

57. Lunn D, Jackson C, Best N, Thomas A, Spiegelhalter D. *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Boca Raton, FL: CRC Press, 2013.

58. Briggs AH, Sculpher M, Claxton K. *Decision Modelling for Health Economic Evaluation*. Oxford: Oxford University Press; 2006.

59. Welton NJ, Sutton AJ, Cooper NJ, Abrams KR, Ades AE. *Evidence Synthesis for Decision Making in Healthcare*. Hoboken, NJ: John Wiley & Sons; 2012. https://doi.org/10.1002/9781119942986

60. Ades AE, Lu G, Claxton K. Expected value of sample information calculations in medical decision modeling. *Med Decis Making* 2004;**24**:207–27. https://doi.org/10.1177/0272989X04263162

61. Lewis G, Pelosi AJ, Araya R, Dunn G. Measuring psychiatric disorder in the community: a standardized assessment for use by lay interviewers. *Psychol Med* 1992;**22**:465–86. https://doi.org/10.1017/S0033291700030415

62. Ritchie J, Spencer L. Qualitative Data Analysis for Applied Policy Research. In Huberman AM, Miles MB, editors. *The Qualitative Researcher's Companion*. Thousand Oaks, CA: Sage Publications, Inc.; 2002. pp. 305–29.

63. Willis GB. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks, CA: Sage Publications, Inc.; 2005.

64. Duffy L, Bacon F, Clarke CS, Donkor Y, Freemantle N, Gilbody S, *et al.* A randomised controlled trial assessing the use of citalopram, sertraline, fluoxetine and mirtazapine in preventing relapse in primary care patients who are taking long-term maintenance antidepressants (ANTLER: ANTidepressants to prevent reLapse in dEpRession): study protocol for a randomised controlled trial. *Trials* 2019;**20**:319. https://doi.org/10.1186/s13063-019-3390-8

# Appendix 1 Minimal clinically important difference on the Beck Depression Inventory, version 2, according to the patient's perspective

See Button *et al.*[36]

# Appendix 2 The relative responsiveness of test instruments can be estimated using a meta-analytic approach: an illustration with treatments for depression

See Kounali *et al.*[37]

# Appendix 3 Using parameter constraints to choose state structures in cost-effectiveness modelling

See Thom *et al.*[38]

# Appendix 4 How much change is enough? Evidence from a longitudinal study on depression in UK primary care

**How much change is enough? Evidence from a longitudinal study on depression in UK primary care.**


**Background**
The Patient Health Questionnaire (PHQ9), the Beck Depression Inventory, 2nd edition (BDI-II) and the Generalised Anxiety Disorder Assessment (GAD-7) are widely used in the evaluation of interventions for depression and anxiety. Little empirical study of the Minimum Clinically Important Difference (MCID) exists for these scales.

**Method**
A prospective cohort of 400 patients in primary care, UK, were interviewed on four occasions, two weeks apart. At each time point, participants completed all three questionnaires and a 'global rating of change' scale (GRS). MCID estimation relied on the reduction in scores in those reporting improvement on the GRS scale. The data was modelled using a Bayesian hierarchical beta-regression stratified by three categories of baseline severity. This method also allowed us to calculate receiver operating characteristics (ROC) parameters.

**Results**
For moderate severity, those who reported improvement had a change of 21% (95% confidence interval (CI) -26.7-14.9) on the PHQ9; 23% (95% CI -27.8 -18.0) on the BDI-II and 26.8% (95% CI -33.5 -20.1) on the GAD-7. Using ROC analysis, the threshold score below which participants were more likely to report improvement than no change were -1.7, -3.5 and -1.5 points on the PHQ9, BDI-II and GAD-7, respectively at moderate severity. This corresponds to 21%, 24% and 27% reduction. At the lowest severity the threshold score rose markedly as a percentage, indicating the difficulty in discriminating change at low severity levels.

**Conclusions**
The self-administered scales had similar characteristics in relation to self-reported improvement. An MCID of about a 20% reduction in scores is a useful rough guide for these scales. The MCID increases, as a percentage, for those at lower severity. This indicates that treatments are unlikely to lead to the experience of benefit in those with low symptoms.


**Keywords:** depression, primary care, BDI-II, PHQ-9, GAD-7, minimal clinically important difference, baseline severity, beta-regression.

## Introduction

Depression is a common reason for consultation in primary care (McManus S *et al.*, 2014) and a major public health problem. Clinicians are faced with the difficulty of making treatment recommendations to patients they see in primary care based upon evidence that used assessments for depressive symptoms that were developed primarily for research purposes. Deciding what constitutes a clinically important treatment effect for those research assessments is therefore essential for interpreting the results of clinical research and designing randomised trials.

The minimum clinically important difference (MCID) provides a measure of the smallest change in an outcome that is perceived as important to patients. The UK National Institute for Health and Care Excellence (NICE) proposed a reduction of three points on the Hamilton Depression Rating Scale as clinically important, but this was based solely on the opinion of an expert group (Kendrick and Pilling, 2012). Others have used approaches that rely upon the error of measurement of scales. (Christensen and Mendoza, 1986, Hays and Hadorn, 1992, Jacobson *et al.*, 1984, Jacobson and Truax, 1991, Kendall PC *et al.*, 1999) but this approach does not incorporate the patients' perspective.

Clinicians and policy makers are giving more emphasis to patients' perspectives in the evaluation of interventions and public health policies. It is therefore important to establish an MCID anchored in the experiences of patients. In previous work, we have investigated the MCID for the Beck Depression Inventory (BDI-II) from the perspective of the patient (Button *et al.*, 2015). Using a Global Rating of Change Scale (GRS), patients were asked whether they felt better, the same, or worse since they were last seen, and the MCID was calculated as the minimum change in depression scores associated with reporting feeling 'better'. This study found that, in absolute terms, the MCID was larger for those with more severe depressive symptoms at baseline, and therefore concluded that MCID might be best conceived as a proportional change (Button *et al.*, 2015). This previous study used data from clinical trials in which patients were only eligible if they exceeded a severity threshold, and thus excluded patients with lower depression scores.

The current study further develops the previous approach. The aim was to estimate the MCID for the BDI-II, PHQ9 and GAD-7 scales. It studies a sample of primary care patients who have been consulting about symptoms of depression and anxiety with broad inclusion criteria to better reflect the population seeking help. We have also extended the work to include the PHQ9 and GAD-7 that are frequently used in research and are the primary outcomes in Improving Access to Psychological Therapies (IAPT) services. The large sample size also allowed us to refine GRS groupings that allow comparisons between those reporting improvement against those reporting "feeling the same" rather than merging the latter group with those "feeling worse". We report on three different approaches to estimate the MCID: the mean change for those "feeling better", the mean difference in change between "feeling better" and "feeling the same", and the threshold value below which participants are more likely to report "feeling better" than report "feeling the same".

## Method

### Participants

The sample was recruited from primary care surgeries in three UK sites (Bristol, Liverpool, and York) between February 2013 and April 2014. This study was part of the PANDA programme (NIHR programme "What are the indications for Prescribing ANtiDepressAnts that will lead to clinical benefit?"; NIHR Programme Grant= RP PG 0610 10048). One of the

primary objectives of this element of the programme was to estimate the MCID for measures of depression by assembling a pragmatic and contemporary cohort of patients seeking help in primary care with a broad range of depression symptom severity. As anxiety symptoms are often co-morbid with depression and no NICE guidelines address such presentations, the study also collected data on a measure of generalised anxiety, the GAD-7, enabling us to additionally explore the MCID for such a measure (Kendrick and Pilling, 2012).

Computerised records at collaborating general practices at each site were searched to identify people who had reported depressive episodes, depressed mood, depressive symptoms or a major depressive episode in the past year. Individuals were included if they were aged between 18 and 74 years, treated or not treated with antidepressants, and referred or not referred to IAPT services. We excluded people who: were diagnosed with bipolar disorder, psychosis or an eating disorder; had alcohol or substance use problems; were unable to complete study questionnaires; or were 30 weeks or more pregnant. Overall, 7,721 patients were sent an information letter in the post 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, and 563 consented to take part in the cohort study. Data on our measures were collected at four time points, each approximately two weeks apart. At time one, 559 people provided data (4 could not be contacted), with corresponding figures at follow-ups two, three and four of 476 (85%), 443 (79%) and 430 (77%) respectively. 400 (72%) participants provided data at each of the four follow-ups and were included in our analyses. Participants missing data at one or more follow-ups were excluded.

Interviews were conducted at the participant's home or GP surgery. 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**

*Beck Depression Inventory–II (BDI-II)*

The BDI-II (Beck *et al.*, 1996) is a self-report measure of the severity of depressive symptoms, consisting of 21 items, each assessed using a 4-point scale ranging from 0 to 3. Possible scores range from 0 to 63. Higher scores indicate a greater severity of depressive symptoms. Participants were asked about the previous 2 weeks.

*Patient Health Questionnaire (PHQ9)*

The PHQ9 (Kroenke and Spitzer, 2002) is a self-report measure of the severity of depressive symptoms, consisting of 9 items each with a 4-point scale ranging from 'Not at all' (0) to 'Nearly every day' (3). Possible scores range from 0 to 27, and higher scores indicate a greater severity of depressive symptoms. The PHQ9 asked about the previous 2 weeks.

*Anxiety*

The Generalised Anxiety Disorder Assessment (GAD-7) (Spitzer *et al.*, 2006) was used to measure anxiety at each time point. The GAD-7 is a self-report measure of generalised anxiety symptoms consisting of 7 items, each assessed using a 4-point scale ranging from 'Not at all' (0) to 'Nearly every day' (3). Possible scores range from 0 to 21. Higher scores indicate a greater severity of anxiety and questions were asked about the previous 2 weeks.

*Global Rating of Change Scale*

The global rating of change scale is a self-report measure of subjective well-being over time, asking participants: "Compared to when we last saw you 2 weeks ago how have your moods and feelings changed?". The five possible responses 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). Participants completed two global rating of change scales (separated by other questionnaires) at each time point, to assess reliability (Kamper *et al.*, 2009, Robinson *et al.*, 2017).

*Clinical Interview Schedule – Revised (CIS-R)*

The CIS-R (Lewis and Pelosi, 1990) is a fully structured self-administered computerised assessment of common mental disorders that has been extensively used in community samples. Participants were assessed using the CIS-R at baseline only. The thresholds used (0-11/12-19/20+) were those pre-specified in the protocol for the subsequent PANDA trial (Salaminios *et al.*, 2017).

*Demographics*

Demographic variables were measured at baseline using a self-administered computerised assessment. These were age, sex, ethnicity, employment status, financial status, and education level.

*Current Antidepressant Use*

A short self-report measure was used to assess current medication use at each time point. Participants were asked whether or not they were currently taking antidepressants.

**Statistical Analyses**

*Accounting for baseline dependency*

We previously found that MCID on the BDI-II in absolute terms varied according to baseline severity, with larger MCID estimates at higher levels of severity (Button *et al.*, 2015). In preliminary analyses in the current study it was also noted that the relationship between the GRS and severity on the three measures was different for participants with low (<=11), medium (12-19) and high (>=20) scores on CIS-R completed at time 1. For example, in Table 1 the average initial PHQ9 score in the group reporting "feeling the same" is lower than in those reporting "feeling better" when baseline severity is low (CIS-R$\leq$11). In contrast, in the high (CIS-R >20) the average initial PHQ9 score was lower in those reporting "feeling better" compared to those reporting "feeling the same". These patterns were similar for all outcomes (Tables 2 and 3). For this reason, we stratified all future analyses according to the three severity groupings and this allowed estimation of group-specific average initial values and differences in change scores across all the time points. Using the CIS-R also conferred the advantage of providing a measure of baseline severity independent of the scales of interest.

*Reliability of the Global Rating of Change Scale (GRS)*

Reliability of the Global Rating of change scale was quantified using the two repeated assessments completed by the patient within each period, in both absolute and relative terms. Absolute levels of agreement were estimated via the (unweighted) Kappa coefficient (Landis and Koch., 1977). We also assessed reliability using the intra-class correlation coefficient (Skrondal and Rabe-Hesketh., 2004). We carried out the calculations using Stata version 15 (StataCorp, 2015).

*Change in BDI-II, PHQ9 and GAD-7 scores - Modelling*

We used Bayesian hierarchical beta-regression models to estimate the changes (as proportions) in symptom scores measured by the three scales (BDI-II and PHQ9 and GAD-7)

and over multiple waves in each of the GRS groupings and baseline CIS-R score (Verkuilen J and M, 2012, Zimprich, 2010). We carried out comparisons of different models using various distributional assumptions and link functions, and found the beta-regression to perform best (Spiegelhalter *et al.*, 2002). We modelled change in symptoms on the proportional scale.

A detailed description of the model specifics, model estimates are provided in the online Appendix 1. We carried out model fitting, model comparisons and post-estimation calculations using the WinBUGS statistical software (Spiegelhalter *et al.*, 2007). Through modelling, we estimated GRS-specific changes over time and potential interactions with the baseline CIS-R. Given the small sample sizes in some GRS response options, these were amalgamated as follows: “I feel a lot better” (1) and “I feel slightly better” (2) under the revised category “Feeling better”; “I feel slightly worse” (4) and “I feel a lot worse” (5) under the revised category: “Feeling worse”.

We express differences in terms of proportional as well as absolute scores using standard post-estimation calculations. The variability in the distribution of change in the different groups was also estimated. The difference in change between the GRS groups in absolute as well as standardised form were also calculated post-estimation to assess the ability of the different instruments to discriminate between the groups.

*Receiver Operating Characteristic (ROC) analysis*

We estimated the threshold value of change that corresponds to the maximum improvement in sensitivity over chance. Estimation of the sensitivity and specificity corresponding to this optimum is a function of the ROC parameters under assumptions of approximate normality (Details in Appendix 1).

The Receiver Operator Characteristic (ROC) parameters required for the derivation of the MCID were based on post-estimation calculations for functions of the parameters of the above regression models. These consist of the standardised difference between the group reporting “feeling better” and the group reporting “feeling the same” as well as the ratio of the variances between the two groups. It should be noted that in previous work (Button *et al.*, 2015) the groups reporting “feeling worse” and “feeling the same” were merged whereas in this work the group reporting “feeling worse” does not contribute any information to the estimation of the threshold value of change which optimally discriminates from the group reporting improvement.

## Results

*Sample Characteristics*

Patients with at least one follow-up visit with data on the GRS was needed to estimate change. 400 patients were included in the analyses and had complete data for all four time points. No baseline differences between excluded and included patients were apparent in the outcomes under study or their demographics. Demographic and clinical characteristics are shown in Table A2.1 (Appendix 2). Participants were aged 17 to 71 years (mean = 48.7), and the majority were female, white, married and employed. Roughly a third of participants had completed higher education. Just under half of participants met ICD-10 criteria for major depressive disorder at baseline. The vast majority reported using antidepressants at each time point.

Descriptive statistics of the distribution of GRS scale over time overall as well as stratified by CIS-R are presented in Appendix 2 (Table A2.2, Figure A2.1). There were no significant changes in GRS scores over time.

*Test-Retest Reliability of the Global Rating Scale*

Absolute levels of agreement were found to be substantial or excellent, with kappa values of 0.73, 0.84, 0.86 and 0.81 for baseline, first, second and third visits respectively. The corresponding levels of agreement were 86%, 90%, 91% and 88% for baseline, first, second and third visits respectively. The intraclass correlation coefficients were: 0.95 (95% CI 0.94, 0.96) at baseline; 0.98 (0.97, 0.99) at the first visit; 0.92 (0.90, 0.94) at the second; and 0.99 (0.98, 0.994) at the third.

*Change in BDI-II, PHQ9 and GAD-7 over time for each grouping of the Global Rating of Change (GRS) scale*

In Table 1 we present estimated mean initial levels and changes in mean scores in both absolute and proportional terms for each CIS-R severity group and GRS group on the PHQ9. Tables 2 and 3 provide the same estimates for the BDI-II and GAD-7 (see methods for an explanation of this analytical approach). The initial scores vary depending upon the CIS-R groups. The changes required for people to report "feeling better" increase with baseline severity (Figures 1-3). It is also noteworthy that the increases seen for those "feeling worse" were not as large as the reductions in those reporting "feeling better".

No differences in the estimated percentage changes for those reporting "feeling better" was found across CIS-R severity groups, for all outcomes (Tables 1-3). In Figures 1-3 we present the changes for those reporting "feeling better" and those reporting "feeling the same" for each of the outcomes as a function of their initial scores.

Participants who reported "feeling the same", also experienced reductions in score on all outcomes. In Table 4 we have estimated the difference in the changes reported by those who report "feeling better" and those who report "feeling the same", in absolute scores as well as a percentage of their respective baseline scores. In general, the differences between "feeling better" and the same became larger as the CIS-R severity increased. For patients with medium levels of CIS-R there was no evidence that these difference in reduction were different to the changes observed for the lower CIS-R category. Only for those with high CIS-R scores at baseline, the difference in reductions between the two groups were significantly larger when compared with lower severity CIS-R groups.

*ROC analysis*

In Table 5 we present our estimates from the ROC analysis. The ROC analysis selects the optimal threshold below which participants are more likely to report "feeling better" rather than "feeling the same". The mean change in the group reporting "feeling better" (see Tables 1-3) is a good approximation for the threshold when the baseline symptom severity is moderate and high for all three instruments. However, when the depression severity is low, the threshold needs to be considerably lower than the mean change in order to optimise the discrimination between the two groups (Figure Appendix 1a-1c).

These results illustrate that at lower levels of depression severity it is much more difficult to discriminate between "feeling better" and "feeling the same" for all three scales. The threshold was estimated at 2 points and was not greatly affected by baseline severity for the PHQ9.

The threshold score for the BDI-II was higher at low baseline severity at 5 points than for moderate and high CIS-R which was 4 points. Finally, the threshold score for GAD-7 was 2 points for low and moderate CIS-R and 1 point for high CIS-R at time 1 (Table 5). What is more important, are the noticeably lower levels of sensitivity of patients' GRS response to identify improvements, when the baseline severity is low. This is true for all measures. At low baseline CIS-R, the sensitivity (Table 5) was 35%, 36% and 32% for PHQ9, BDI-II and GAD-7, respectively, indicating the proportion who reported they felt better and had experienced reductions larger than the threshold score. At higher baseline CIS-R, the patients who reported improvement had much higher chances (60% or more) to show reductions larger than the threshold score in all scales.

It should also be noted that there is uncertainty in the presented values of the optimal thresholds. These uncertainties are as large as the differences between these values across CIS-R groups. Thus, we do not have evidence that the threshold scores vary according to severity. However, this implies that the threshold as a percentage reduction is increasing as the severity drops (Table 5). Uncertainty estimates of the sensitivity and specificity at the optimal threshold are also presented in Table SA1.1 (Appendix 1). Statistics relevant to the determination of the optimal threshold and effect size calculations, namely: standard deviations of baseline scores and changes scores are also presented in Table SA1.2 (Appendix 1).

## Discussion

We have estimated the minimally clinically important difference using a patient-centred approach for three commonly used scales used to assess depression and anxiety. We have estimated the reduction in scores during the previous 2 weeks in those who reported "feeling better". We then estimated the difference between "feeling better" and "feeling the same" in terms of the reduction of scores.

The finding that people who reported "feeling the same" also had a small reduction in symptoms is not well understood (Robinson *et al.*, 2017). The patients' GRC is likely to include constructs additional to those measured by the disease specific scales, so a perfect correlation is not expected. Research in health related quality of life have also found that retrospective measures of the patient's view of change is sensitive to change in disease-specific scales and correlates strongly with patient's satisfaction with change but is not concordant with repeated current assessments of patients' experience of change (Fischer *et al.*, 1999). This literature, also presents evidence that those with less severe dysfunction at baseline have smaller change score over time, thus, variability on baseline dysfunction may also reduce the strength of association between change scores and the GRC (Stucki *et al.*, 1996). The reductions we observed in this study was proportionally more dramatic amongst those with lower severity.

Finally, we also formulated the problem as trying to distinguish between "feeling better" and "feeling the same" using ROC analysis to estimate the optimal threshold to provide separation. This final method seems the most robust as it can take account of the increased variability of scores at the lower severity.

In the lowest severity group, average reductions experienced by those reporting "feeling better" were estimated at 24.1%, 30.8% and 26.4% on the PHQ9 and BDI-II and GAD-7 scales respectively. However, the optimal threshold required to discriminate between "feeling better" and "feeling the same" were reductions of 48%, 51.5% and 71% respectively. The thresholds at the middle level severity were 21%, 23% and 26.8% respectively.

The marked increase of threshold in percentage terms is because the variability, particularly in those "feeling better", is relatively large in those at lower severity so this makes discrimination more difficult.

In our previous work we found evidence that viewing the MCID as a proportion led to a more constant value over the severity range (Button *et al.*, 2015). However, this was based on analyses informed by RCTs which excluded patients below a certain threshold score and similar distributions of baseline scores on the BDI scale. In this study with a sample with lower severity scores, it is apparent that there is still an increase in MCID in proportional terms at lower levels of severity, even if the absolute levels are relatively constant. It is perhaps unsurprising that those with low scores will find it more difficult to distinguish between "feeling the same" and "feeling better". These results bring to foreground the concept of reliability of change in outcome scales and its dependence with baseline severity. There it seems that baseline scores below certain thresholds render the quantification of change in proportionate terms less informative with respect to patients' retrospective evaluations.

The use of ROC analysis also allowed the evaluation of performance of the ability of patients' GRS scoring to identify change in outcomes frequently used in RCTs, at the threshold score, namely overall discrimination (AUC) and sensitivity and specificity (Table 5, Table SA1.1). Only a small proportion of people reporting improvement at low baseline severity actually show reductions larger than the threshold score, in all scales (35%;36%; 32% for PHQ9, BDI-II and GAD-7, Table 5). This proportion is also significantly lower compared to the rest of the CIS-R groupings, for all three outcome measures (Table SA1.1). This implies that even if treatment effects are similar in those with less severe symptoms, it is much less likely that they will experience any benefit. This confirms that knowledge about treatment effects and the MCID should allow, in principle, to determine whether an individual is likely to benefit from a treatment.

This is the first large cohort study in primary care exploring this question and to our knowledge, there is only one study exploring a similar question and reached similar conclusions. This study used data from a small RCT and explored the question of the size of effect that could be considered as a successful treatment outcome (McMillan *et al.*, 2010), based on the reliable and clinically significant change (RCSC) index and using the PHQ9. The reported proportions of patients experiencing improvements was significantly reduced among asymptomatic patients (PHQ9$\leq$4) and found that the odds of improvement could be affected by how the RCSC index was anchored e.g. how reliably patients' change could be discriminated against a clinical mean rather a non-clinical one.

It is striking that there are many similarities in how the different scales behave in relation to self-reported improvement. Previous meta-analytic work evaluating the relative responsiveness of eight scales (6 depression and 2 quality of life) also found little difference between scales capturing change caused by treatment (Kounali *et al.*, 2016). That study included a broad range of different treatments from RCTs and even though the absolute values of the scales differed, the pattern of results was similar and the proportionate changes seemed comparable.

*Strength and limitations*

This is the first study of a large contemporary cohort drawn from a population seeking help for their symptoms in primary care in the UK. In contrast to our previous study that used data from RCTs, this sample was not selected according to severity criteria so included less severe patients and also minimised any regression to the mean. We used a flexible

Bayesian approach towards estimation and were careful in ensuring the robustness of our statistical models. In particular, our approach provided a realistic assessment of the distribution of change, which is critical for the determination of the optimal threshold through ROC analyses. These results enhance our earlier work by extending it to lower severities of symptoms and to include other commonly used outcome measures, the PHQ9 and GAD-7.

Despite the size of this cohort, the number with low CIS-R baseline severity who report "feeling better" at baseline is still rather small (n=36), so some of our estimates lacked precision. Our method also relied on the use of self-reported improvement. It remains unclear how patients' perceptions of change can inform therapeutic significance, but it is certainly an aspect of this. It is also noteworthy that those who reported "feeling the same" experienced a reduction in symptoms, and there was a marked asymmetry in this sample such that feeling worse was not associated with such large changes as "feeling better". The reasons for this are unknown. In our analyses we could take account of the changes in those "feeling the same" when estimating the MCID. Using self-reported change as a "gold standard" has good face validity (Malpass *et al.*, 2016) and qualitative findings support its use. Yet our results indicate areas where our understanding of the responses requires further research.

*Implications*

Our results have three potential uses. Firstly, they have implications for sample size calculations for RCTs using these outcomes. The MCID estimates can be used as a basis for sample size estimation if the likely values of the outcome at follow-up are known given that the MCID varies according to severity, at least in proportional terms. Our best estimates are the initial values given in Table 5. However, the application is not straightforward. Here we have estimated an average within-person change related to improvement but an RCT compares groups. Application of our results would require a counterfactual argument in which researchers compare the same individual(s) who receive placebo but who might have received the active treatment. The MCID estimated from a within person calculation can then be applied to the between group differences in a clinical trial.

The MCID estimates could ultimately guide decisions about whether a treatment will benefit an individual. For this, one needs to be able to predict the likely score for that person on the proposed outcome measure were they not to receive that treatment. This is available within an RCT design since treated and control patients are exchangeable and thus control subject scores at follow-up provide us with a good guess on a patient's potential outcome at follow-up. We then compare the treated individual's attained score at follow-up with the likely scores attained by the controls, to see if the likely treatment benefits exceed the MCID. Our results indicate that even if treatment effects are similar in those with less severe symptoms, it is much less likely that they will experience any benefit.

The third application is in interpreting the results of clinical trials. Using a similar argument, the MCID could be used to decide whether patients would experience a clinically meaningful benefit from the treatment when the treatment effect is larger than the MCID. Characterisation of the profile of treated patients who experienced reductions larger than the MCID could also be useful.

There is currently much controversy about the benefits or otherwise of antidepressant treatment, especially in those with less severe symptoms. We regard our approach here as a step towards resolving this controversy using empirical data. In order for us to be confident about recommending treatments to patients we will need more accurate information on individualised treatment effects, the outcome without treatment as well as the MCID.

**Acknowledgements**

This paper is independent research funded by the National Institute for Health Research (Programme Grants for Applied Research, What are the indications for Prescribing ANtiDepressAnts that will lead to a clinical benefit: PANDA RP-PG-0610-10048). This study was also supported by the NIHR Biomedical Research Centre at the University Hospitals Bristol NHS Foundation Trust and the University of Bristol. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. We are grateful to all the patients, practitioners and GP surgery staff who took part in PANDA. We also thank those colleagues who contributed to the PANDA study, through recruitment and retention of patients, provision of administrative support, or delivery/supervision of therapy. Finally, we are grateful to all our colleagues who were involved with the studies as co-applicants but who have not participated in drafting this manuscript.

## References

**Beck, A., Steer, R. & Brown, G.** (1996). Beck Depression Inventory-II. San Antonio. In *TX: Harcourt Brace & Company*.

**Button, K., Kounali, D., Thomas, L., Wiles, N., Peters, T., Welton, N., Ades, A. & Lewis, G.** (2015). Minimal clinically important difference on the Beck Depression Inventory-II according to the patient's perspective. *Psychological medicine* **45**, 3269-3279.

**Christensen, L. & Mendoza, J.** (1986). A method of assessing change in a single subject: An alteration of the RC index. *Behavior Therapy* **17**, 305-308.

**Fischer, D., Stewart, A. L., Bloch, D. A., Lorig, K., Laurent, D. & Holman, H.** (1999). Capturing the patient's view of change as a clinical outcome measure. *Jama* **282**, 1157-62.

**Hays, R. D. & Hadorn, D.** (1992). Responsiveness to change: an aspect of validity, not a separate dimension. *Qual Life Res* **1**, 73-75.

**Jacobson, N. S., Follette, W. C. & Revenstorf, D.** (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy* **15**, 336-352.

**Jacobson, N. S. & Truax, P.** (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol* **59**, 12-9.

**Kamper, S. J., Maher, C. G. & Mackay, G.** (2009). Global Rating of Change Scales: A Review of Strengths and Weaknesses and Considerations for Design. *The Journal of Manual & Manipulative Therapy* **17**, 163-170.

**Kendall PC, Marrs-Garcia A, Nath SR & RC., S.** (1999). Normative comparisons for the evaluation of clinical change Journal of consulting and clinical psychology. *J Consult Clin Psychol.* **67**, 285-99.

**Kendrick, T. & Pilling, S.** (2012). Common mental health disorders — identification and pathways to care: NICE clinical guideline. *The British Journal of General Practice* **62**, 47-49.

**Kounali, D. Z., Button, K. S., Lewis, G. & Ades, A. E.** (2016). The relative responsiveness of test instruments can be estimated using a meta-analytic approach: an illustration with treatments for depression. *J Clin Epidemiol* **77**, 68-77.

**Kroenke, K. & Spitzer, R. L.** (2002). The PHQ-9: a new depression diagnostic and severity measure. *Psychiatric annals* **32**, 509-515.

**Landis, J. R. & Koch., G. G.** (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**.

**Lewis, G. & Pelosi, A.** (1990). Manual of the revised clinical interview schedule (CIS-R). *Institute of Psychiatry, London*.
**Malpass, A., Dowrick, C., Gilbody, S., Robinson, J., Wiles, N., Duffy, L. & Lewis, G.** (2016). Usefulness of PHQ-9 in primary care to determine meaningful symptoms of low mood: a qualitative study. *The British Journal of General Practice* **66**, e78-e84.
**McManus S, Meltzer H, Brugha T, Bebbington P & R., J.** (2014). Adult psychiatric morbidity in England 2007: results of a household survey. pp. 1-27. NHS Information Centre: Leeds.
**McMillan, D., Gilbody, S. & Richards, D.** (2010). Defining successful treatment outcome in depression using the PHQ-9: A comparison of methods. *Journal of Affective Disorders* **127**, 122-129.
**Robinson, J., Khan, N., Fusco, L., Malpass, A., Lewis, G. & Dowrick, C.** (2017). 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**.
**Salaminios, G., Duffy, L., Ades, A., Araya, R., Button, K. S., Churchill, R., Croudace, T., Derrick, C., Dixon, P., Dowrick, C., Gilbody, S., Hollingworth, W., Jones, V., Kendrick, T., Kessler, D., Kounali, D., Lanham, P., Malpass, A., Peters, T. J., Riozzie, D., Robinson, J., Sharp, D., Thomas, L., Welton, N. J., Wiles, N. & Lewis, G.** (2017). 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 a randomised controlled trial. *Trials* **18**, 496.
**Skrondal, A. & Rabe-Hesketh., S.** (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC: Boca Raton, FL.
**Spiegelhalter, D., Thomas, A., Best, N. & Lunn, D.** (2007). WinBUGS User Manual Version 1.4 January 2003. Upgraded to Version 1.4.3.
**Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van Der Linde, A.** (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**, 583-639.
**Spitzer, R. L., Kroenke, K., Williams, J. B. & Löwe, B.** (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of internal medicine* **166**, 1092-1097.
**StataCorp** (2015). Stata Statistical Software: Release 14. . StataCorp LP: College Station, TX.
**Stucki, G., Daltroy, L., Katz, J. N., Johannesson, M. & Liang, M. H.** (1996). Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. *J Clin Epidemiol* **49**, 711-7.
**Verkuilen J & M, S.** (2012). Mixed and mixture regression models for continuous bounded responses using the beta distribution. *J Educ Behav Stat* **37**, 82-113.
**Zimprich, D.** (2010). Modeling change in skewed variables using mixed beta regression models. *Res Hum Dev* **7**, 9-26.

**Table 1: Estimate initial and change in PHQ9 score (previous 2 weeks) according to patient reported Global ratings and time 1 CIS-R**

| Global Rating Scale | Feeling Better | | | Feeling Same | | | Feeling Worse | | |
|---|---|---|---|---|---|---|---|---|---|
| | Mean | 95% CI | | Mean | 95% CI | | Mean | 95% CI | |
| Baseline CIS-R | **Initial PHQ9** | | | | | | | | |
| ≤11 | 4.15 | [3.07 | 5.39] | 2.66 | [2.15 | 3.26] | 6.08 | [3.38 | 9.59] |
| 12-19 | 7.97 | [6.08 | 10.17] | 8.75 | [7.51 | 10.11] | 11.06 | [7.49 | 15.25] |
| 20+ | 12.20 | [9.91 | 14.60] | 15.01 | [13.98 | 16.07] | 17.23 | [15.11 | 19.09] |
| | **Change in previous 2 weeks** | | | | | | | | |
| ≤11 | -1.00 | [-1.45 | -0.63] | -0.28 | [-0.48 | -0.09] | 0.745 | [0.15 | 1.41] |
| 12-19 | -1.66 | [-2.28 | -1.11] | -0.83 | [-1.33 | -0.32] | 0.447 | [-0.20 | 1.12] |
| 20+ | -2.38 | [-2.85 | -1.88] | -0.66 | [-1.05 | -0.26] | 0.270 | [-0.15 | 0.72] |
| | **Change in previous weeks as a proportion of initial PHQ9 score** | | | | | | | | |
| ≤11 | -0.24 | [-0.31 | -0.17] | -0.10 | [-0.17 | -0.03] | 0.13 | [0.03 | 0.24] |
| 12-19 | -0.21 | [-0.27 | -0.15] | -0.09 | [-0.15 | -0.04] | 0.04 | [-0.02 | 0.10] |
| 20+ | -0.20 | [-0.24 | -0.15] | -0.04 | [-0.07 | -0.02] | 0.02 | [-0.01 | 0.04] |

**Table 2: Estimate initial and change in BDI-II score (previous 2 weeks) according to patient reported Global ratings and time 1 CIS-R**

| Global Rating Scale | Feeling Better | | | Feeling Same | | | Feeling Worse | | |
|---|---|---|---|---|---|---|---|---|---|
| | Mean | 95% CI | | Mean | 95% CI | | Mean | 95% CI | |
| Baseline CIS-R | **Initial BDI** | | | | | | | | |
| ≤11 | 9.67 | [7.52 | 12.02] | 6.24 | [5.29 | 7.27] | 11.74 | [7.07 | 17.37] |
| 12-19 | 14.68 | [11.32 | 18.78] | 15.94 | [13.69 | 18.26] | 16.54 | [10.99 | 22.76] |
| 20+ | 22.25 | [18.80 | 26.07] | 26.99 | [24.77 | 29.13] | 31.68 | [28.05 | 35.50] |
| | **Change per 2 weeks** | | | | | | | | |
| ≤11 | -2.97 | [-3.89 | -2.19] | -1.28 | [-1.67 | -0.93] | 0.11 | [-0.79 | 1.02] |
| 12-19 | -3.36 | [-4.46 | -2.49] | -1.61 | [-2.31 | -0.93] | -0.12 | [-1.00 | 0.78] |
| 20+ | -4.32 | [-5.16 | -3.51] | -1.57 | [-2.20 | -0.94] | 0.14 | [-0.61 | 0.93] |
| | **Change per 2 weeks as a percentage of initial BDI score** | | | | | | | | |
| ≤11 | -0.31 | [-0.36 | -0.25] | -0.21 | [-0.25 | -0.16] | 0.01 | [-0.07 | 0.09] |
| 12-19 | -0.23 | [-0.28 | -0.18] | -0.10 | [-0.15 | -0.06] | -0.01 | [-0.06 | 0.05] |
| 20+ | -0.20 | [-0.23 | -0.16] | -0.06 | [-0.08 | -0.03] | 0.01 | [-0.02 | 0.03] |

**Table 3: Estimate initial and change in GAD-7 score (previous 2 weeks) according to patient reported Global ratings and time 1 CISR**

| Global Rating Scale | Feeling Better | | | Feeling Same | | | Feeling Worse | | |
|---|---|---|---|---|---|---|---|---|---|
| | Mean | 95% CI | | Mean | 95% CI | | Mean | 95% CI | |
| Baseline CIS-R | **Initial GAD-7** | | | | | | | | |
| ≤11 | 3.05 | [2.26 | 4.01] | 1.92 | [1.51 | 2.37] | 5.24 | [3.07 | 7.88] |
| 12-19 | 5.62 | [4.14 | 7.26] | 6.23 | [5.25 | 7.33] | 5.24 | [3.02 | 7.82] |
| 20+ | 9.02 | [7.37 | 10.93] | 11.12 | [10.18 | 11.98] | 13.97 | [12.38 | 15.39] |
| | **Change per 2 weeks** | | | | | | | | |
| ≤11 | -0.81 | [-1.18 | -0.50] | -0.27 | [-0.44 | -0.11] | 0.86 | [0.27 | 1.54] |
| 12-19 | -1.50 | [-2.04 | -1.03] | -0.53 | [-0.92 | -0.11] | 0.00 | [-0.47 | 0.50] |
| 20+ | -1.56 | [-1.96 | -1.16] | -0.42 | [-0.76 | -0.08] | 0.67 | [0.28 | 1.06] |
| | **Change per 2 weeks As a percentage of baseline GAD7 score** | | | | | | | | |
| ≤11 | -0.26 | [-0.34 | -0.19] | -0.14 | [-0.21 | -0.06] | 0.17 | [0.05 | 0.30] |
| 12-19 | -0.27 | [-0.34 | -0.20] | -0.09 | [-0.15 | -0.02] | 0.00 | [-0.09 | 0.10] |
| 20+ | -0.17 | [-0.22 | -0.13] | -0.04 | [-0.07 | -0.01] | 0.05 | [0.02 | 0.08] |

**Table 4: Estimated difference in change between the group reporting feeling better and the group reporting feeling the same in absolute scores and % from their respective initial scores for PHQ9, BDI-II and GAD-7 scales**

| Baseline Severity | CIS-R 0-11 | | | CIS-R 12-19 | | | CIS-R 20+ | | |
|---|---|---|---|---|---|---|---|---|---|
| | Mean | 2.50% | 97.50% | Mean | 2.50% | 97.50% | Mean | 2.50% | 97.50% |
| **Outcome** | **Difference in Change** | | | | | | | | |
| **PHQ9** | -0.73 | -1.13 | -0.40 | -0.85 | -1.45 | -0.31 | -1.70 | -2.18 | -1.24 |
| **BDI** | -1.66 | -2.54 | -0.89 | -1.76 | -2.74 | -0.79 | -2.77 | -3.61 | -1.94 |
| **GAD-7** | -0.54 | -0.88 | -0.24 | -0.99 | -1.53 | -0.49 | -1.15 | -1.57 | -0.72 |
| | **Difference in % Change** | | | | | | | | |
| **PHQ9** | -0.14 | -0.22 | -0.07 | -0.12 | -0.18 | -0.06 | -0.15 | -0.19 | -0.11 |
| **BDI** | -0.10 | -0.16 | -0.04 | -0.13 | -0.18 | -0.08 | -0.14 | -0.17 | -0.10 |
| **GAD-7** | -0.12 | -0.21 | -0.04 | -0.18 | -0.26 | -0.11 | -0.14 | -0.18 | -0.09 |

**Table 5: Estimated threshold score for discriminating between feeling better and feeling the same for the PHQ9, BDI-II and GAD-7 scales, according to baseline severity and related ROC parameters.**

| Instrument Scale | Severity | Threshold score | Threshold as % of baseline | 95% change as % of baseline | Spec [2] | Sens [3] | AUC [4] Mean | [2.5% | 97.5%] |
|---|---|---|---|---|---|---|---|---|---|
| **PHQ9** | ≤11 | -2.0 | 48.2 | [65.1 37.1] | 0.78 | 0.35 | 0.57 | [0.54 | 0.60] |
| | 12-19 | -1.7 | 21.3 | [27.9 16.7] | 0.59 | 0.51 | 0.57 | [0.53 | 0.62] |
| | 20+ | -2.4 | 19.7 | [24.2 16.4] | 0.63 | 0.52 | 0.61 | [0.58 | 0.64] |
| **BDI** | ≤11 | -5.0 | 51.7 | [66.6 41.6] | 0.80 | 0.36 | 0.59 | [0.55 | 0.63] |
| | 12-19 | -3.5 | 23.8 | [30.9 18.6] | 0.65 | 0.50 | 0.60 | [0.55 | 0.65] |
| | 20+ | -4.4 | 19.7 | [23.4 16.9] | 0.65 | 0.51 | 0.61 | [0.58 | 0.65] |
| **GAD-7** | ≤11 | -2.2 | 72.1 | [97.3 54.8] | 0.78 | 0.32 | 0.55 | [0.53 | 0.59] |
| | 12-19 | -1.5 | 26.7 | [36.2 20.7] | 0.51 | 0.62 | 0.59 | [0.55 | 0.64] |
| | 20+ | -0.8 | 8.9 | [10.9 7.3] | 0.55 | 0.54 | 0.57 | [0.54 | 0.59] |

**[1] SD Feeling Same/SD Feeling Better**

**[2] Probability (Improvements/reductions smaller than MCID when Feeling the Same)**

**[3] Probability (Improvements/reductions larger than MCID when Feeling Better)**

**[4] Probability the improvement (reduction) in scores for a randomly chosen patient drawn from those reporting feeling the same is smaller than for a randomly chosen person drawn from those reporting feeling better**

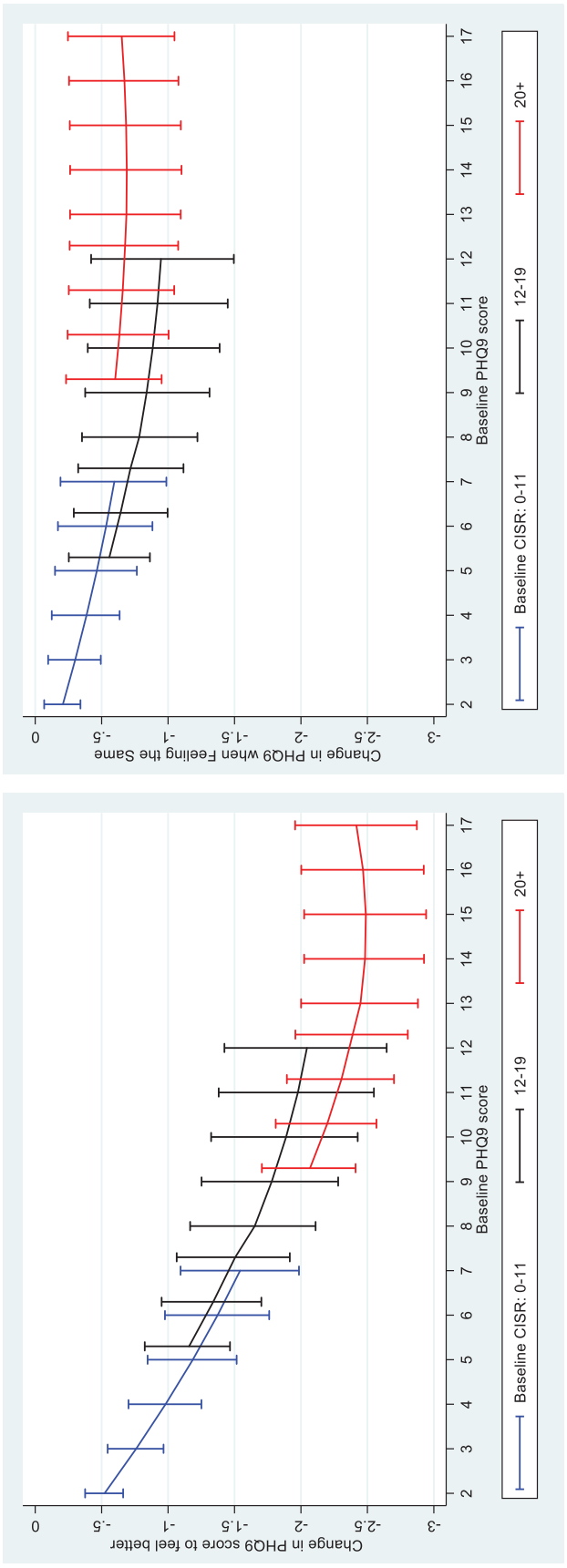
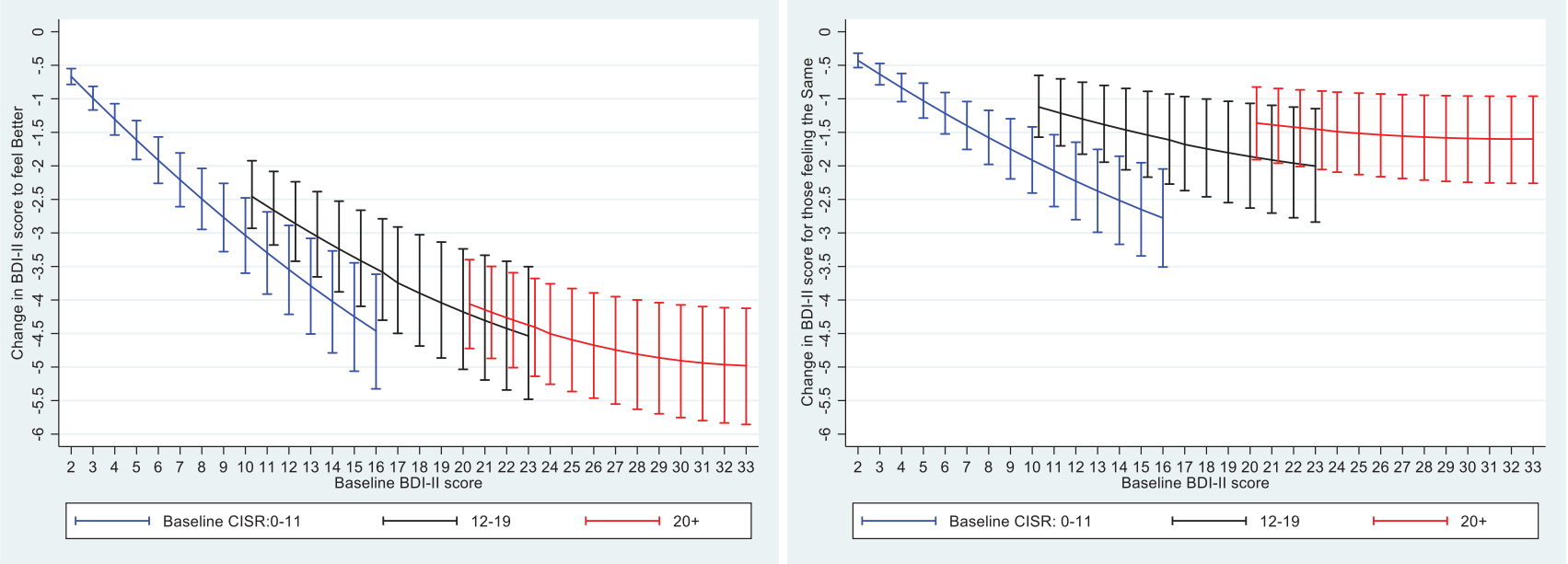



*Figure 1a: Plots of estimated average absolute change in scores of those reporting feeling better (left panel) and those reporting feeling the same (right panel) for PHQ-9 according to their baseline scores controlling for baseline symptom severity assessed by the CIS-R.*

*Figure 1b: Plots of estimated average absolute change in scores of those reporting feeling better (left panel) and those reporting feeling the same (right panel) for BDI-II according to their baseline scores controlling for baseline symptom severity assessed by the CIS-R.*



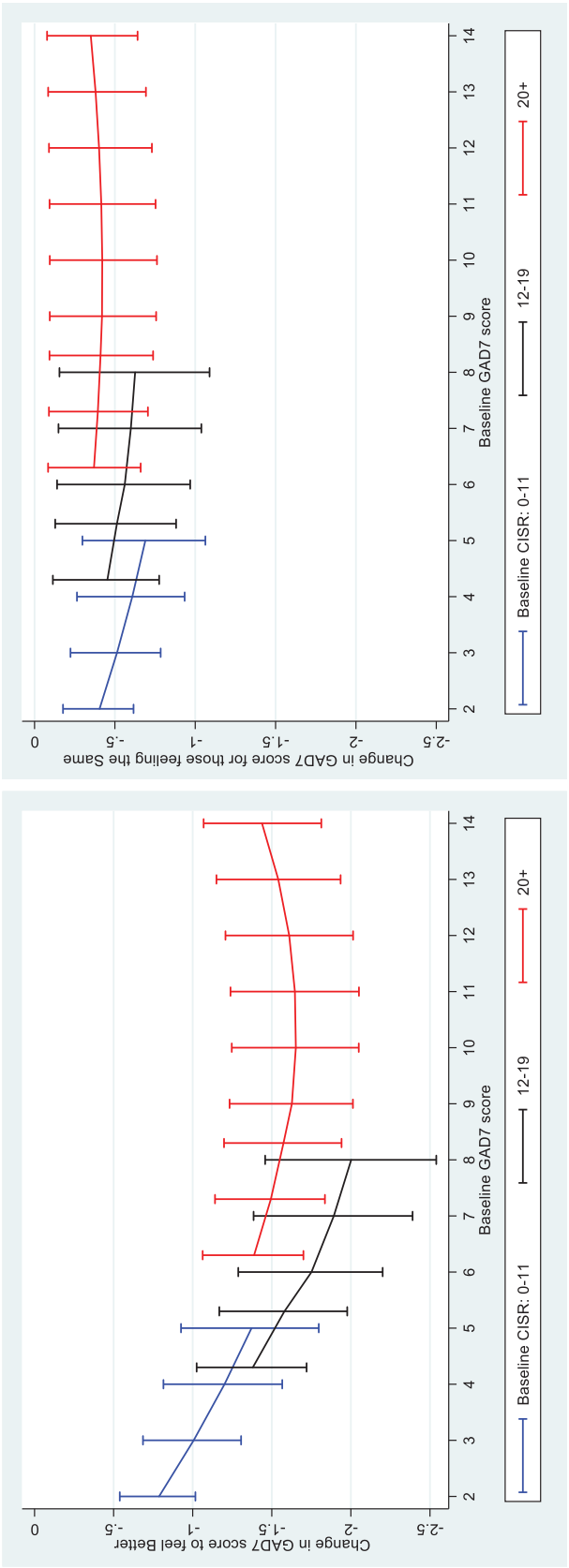



*Figure 1b: Plots of estimated average absolute change in scores of those reporting feeling better (left panel) and those reporting feeling the same (right panel) for GAD-7 (panel c) according to their baseline scores controlling for baseline symptom severity assessed by the CIS-R.*

## Appendix 1
## Beta Regression Model

Beta regression modelling can have substantial advantages when outcomes are bounded and exhibit high levels of skewness. These include substantial improvements in fit as well as increased precision for individual predictions. Beta regression also models outcomes on a multiplicative scale.

More importantly, beta regression allows us to simultaneously explore covariate effects not only on our expectations but also variability which is important for the receiver operating curve (ROC parameters) estimation. The quantification of variability is very often un-appreciated and selectively reported if at all. This state of affairs is despite its importance in sample size calculations required for the design of RCTs as well as meta-analytic studies. Recent methodological advances have allowed a more widespread use of generalised location/scale modelling such as mixed effects beta regression through standard statistical software and for more complex settings e.g. repeated measures analyses.

There is a difference between the regression model we used and a binary regression models where the outcome is whether the patient reports an improvement as a function of the change in their depression scores. The former regression model assumes that the expected value of outcome score change depends on the patient's view of their condition whereas the later assumes that the expected value of the patient's view of how they feel depends on their change in scores of BDI. For this reason, we based the estimation of the required ROC parameters on the regression model we described in the previous section.

Each of the outcomes Y=PHQ9, BDI-II and GAD-7 all of which are bounded within (a,b), where a=0 and b=27 for PHQ9; b=63 for BDI-II and b=21 for GAD-7
We transformed the scale to (0,1) interval by applying the transformation $Y_{new}=(Y-a)/(b-a)$
The reparameterization used for modelling the mean and variance of the beta distribution follows Ferrari and Cribari-Neto (2004), had already appeared in the literature, for example in Jorgensen (1997) or in Cepeda (2001).

$$Y_{(new)ij} \sim Beta(\phi_{ij}\mu_{ij}, \phi_{ij}(1-\mu_{ij}))$$

where i indexes individuals and j indexes visits with j=1,2,3,4

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \alpha_{i,CISR_i,GRS_i} + \beta_{i,CISR_i,GRS_i} * (j-1)$$

$$\phi_{ij} = \exp(-\delta_0 - \delta_1 * (j-1) - \delta_{2,CISR_i} - \delta_{3,GRS_i})$$

Where $\mu_{ij}$ is the conditional mean and $\phi_{ij}$ can be interpreted as a precision parameters, in the sense that for fixed values of the mean $\mu_{ij}$ larger value of $\phi_{ij}$ correspond to larger values of the variance for the outcome $Y_{(new)ij}$

The variance is given by:

$$\sigma_{ij}^2 = \frac{\mu_{ij}(1-\mu_{ij})}{\phi_{ij}+1}$$

The parameter $\beta_{i,CISR_i,GRS_i}$ represent the change between successive visits and is specific to each CIS-R group and GRS group (log-odds for increase in the outcome). Similarly the parameter $\alpha_{i,CISR_i,GRS_i}$ represents the baseline values and is specific to each CIS-R group and GRS group

Note, that both intercepts and slopes in this models are also indexed by individuals and these are assumed to be jointly distributed as bivariate Normal distribution with mean zero and a 2X2 variance-covariance, that is also estimated.

Below, we provide the derivations used to compute change on the original scale and percentage change for each group as a function of the model parameters.

Denoting odds parameter (:=exp( $\beta_{i,CISR_i,GRS_i}$ ) as $\lambda_i, i=1,2,3$ denoting the groups reporting feeling better, same or worse respectively

Then the translation to change on the original scale and proportionate change relative to the groups baseline $p_i$ is a function the estimated odds and the baseline as follows:

$$\frac{p_i'}{1-p_i'} = \lambda_i \frac{p_i}{1-p_i} \text{ then } \frac{p_i'}{p_i} = \frac{\lambda_i}{1+(\lambda_i-1)p_i}$$

$$\text{Change: } p_i' - p_i = \frac{p_i(1-p_i)(\lambda_i-1)}{1+(\lambda_i-1)p_i}$$

$$\text{or \% change } \frac{p_i'}{p_i} - 1 = \frac{(\lambda_i-1)(1-p_i)}{1+(\lambda_i-1)p_i}$$

where $p_i$ are the outcome values on the transformed ([0-1]) scale.

**MCID determination**

Let $\mu_{1,X}$ and $\mu_{0,X}$ denote the mean of the diagnostic outcome: BDI change (log-ratio) at the gold standard disease status (not feeling better) and feeling better respectively and additional covariates X.

$\sigma_1^2, \sigma_2^2$ denote the variances of the outcome for the non-diseased (those feeling better) and diseased groups.

The ROC parameters are:

$$\alpha_X = \frac{(\mu_{1,X} - \mu_{o,X})}{\sigma_1}$$

and $\beta = \frac{\sigma_2}{\sigma_1}$

Then the Area under the curve A is

$$A = \Phi\left(\frac{\alpha_X}{\sqrt{1+\beta^2}}\right)$$

The area under the curve is equal to the probability that the outcome for a randomly drawn diseased subject is higher than the randomly drawn non-diseased individual. ($\Phi(\quad)$: represents the standard cumulative normal density)

The sensitivity at given specificity is : Fo any given (1-specificity), p, the underlying sensitivity is:

$$q(p) = \Phi(\alpha_X + \beta\Phi^{-1}(p))$$

Finally the Maximum improvement of sensitivity over chance (Youden index, Figure A1): This is the maximum difference in observed sensitivity and sensitivity at change (lying on a $45^0$ line in ROC space) over all values of specificity. The corresponding (1-specificity) denoted by $p_{YOUDEN}$ is given by:

$$p_{Youden} = \Phi\{\frac{\left[\left[-\alpha_X\beta + (\alpha_x^2 + 2(\beta^2-1)\log(\beta))^{\frac{1}{2}}\right]\right]}{\beta^2 - 1}\}$$

**Table SA1.1: Uncertainty estimates surrounding the ROC performance characteristics of the MCID on the PHQ9, BDI-II and GAD-7**

| | | Specificity | | | Sensitivity | | |
|---|---|---|---|---|---|---|---|
| Outcome | Baseline CIS-R | Median | 95% CI | | Median | 95% CI | |
| PHQ9 | 0-11 | 0.78 | 0.61 | 0.85 | 0.35 | 0.27 | 0.49 |
| | 12-19 | 0.58 | 0.32 | 0.75 | 0.54 | 0.35 | 0.76 |
| | 20+ | 0.63 | 0.49 | 0.73 | 0.53 | 0.42 | 0.65 |
| BDI-II | 0-11 | 0.80 | 0.69 | 0.85 | 0.36 | 0.30 | 0.45 |
| | 12-19 | 0.65 | 0.42 | 0.77 | 0.50 | 0.37 | 0.69 |
| | 20+ | 0.65 | 0.50 | 0.75 | 0.51 | 0.41 | 0.65 |
| GAD-7 | 0-11 | 0.79 | 0.54 | 0.86 | 0.33 | 0.25 | 0.53 |
| | 12-19 | 0.51 | 0.33 | 0.67 | 0.63 | 0.45 | 0.78 |
| | 20+ | 0.56 | 0.37 | 0.72 | 0.54 | 0.37 | 0.73 |

In Table SA1.3, we depict the uncertainty surrounding the ROC performance characteristics of the MCID and considerable uncertainty is apparent. This uncertainty when propagated leads to uncertainty for the optimal threshold. This is considerable relative to the apparent differences between MCID estimates at different baseline severity levels.

**Table SA1.2: Baseline SD and SD estimates for change on the PHQ9, BDI-II and GAD-7 for the group reporting feeling better**

| | | Feeling Better | | | | Feeling The Same | | | |
|---|---|---|---|---|---|---|---|---|---|
| Outcome | Baseline CIS-R | SD Baseline | 95% CI | SD Change | 95% CI | SD Baseline | 95% CI | SD Change | 95% CI |
| PHQ9 | 0-11 | 2.60 | [2.20 3.06] | 3.35 | [2.86 3.89] | 2.52 | [2.16 2.93] | 3.24 | [2.81 3.70] |
| | 12-19 | 2.08 | [1.79 2.43] | 2.68 | [2.35 3.07] | 2.50 | [2.18 2.86] | 3.21 | [2.86 3.60] |
| | 20+ | 3.60 | [2.91 4.24] | 4.64 | [3.80 5.40] | 3.01 | [2.66 3.40] | 3.87 | [3.50 4.25] |
| BDI-II | 0-11 | 4.92 | [4.04 5.90] | 6.21 | [5.21 7.31] | 4.05 | [3.40 5.09] | 5.08 | [4.39 5.85] |
| | 12-19 | 3.69 | [3.15 4.25] | 4.64 | [4.04 5.26] | 3.76 | [3.24 4.72] | 4.71 | [4.12 5.39] |
| | 20+ | 6.09 | [4.88 7.42] | 7.68 | [6.24 9.15] | 4.50 | [3.76 5.65] | 5.64 | [4.82 6.53] |
| GAD-7 | 0-11 | 2.33 | [1.90 2.82] | 3.04 | [2.51 3.64] | 2.33 | [1.97 2.71] | 3.02 | [2.61 3.47] |
| | 12-19 | 1.90 | [1.63 2.21] | 2.48 | [2.15 2.84] | 2.45 | [2.14 2.80] | 3.18 | [2.81 3.60] |
| | 20+ | 3.43 | [2.76 4.03] | 4.48 | [3.69 5.14] | 2.54 | [2.06 2.99] | 3.30 | [2.71 3.84] |

# Appendix 2

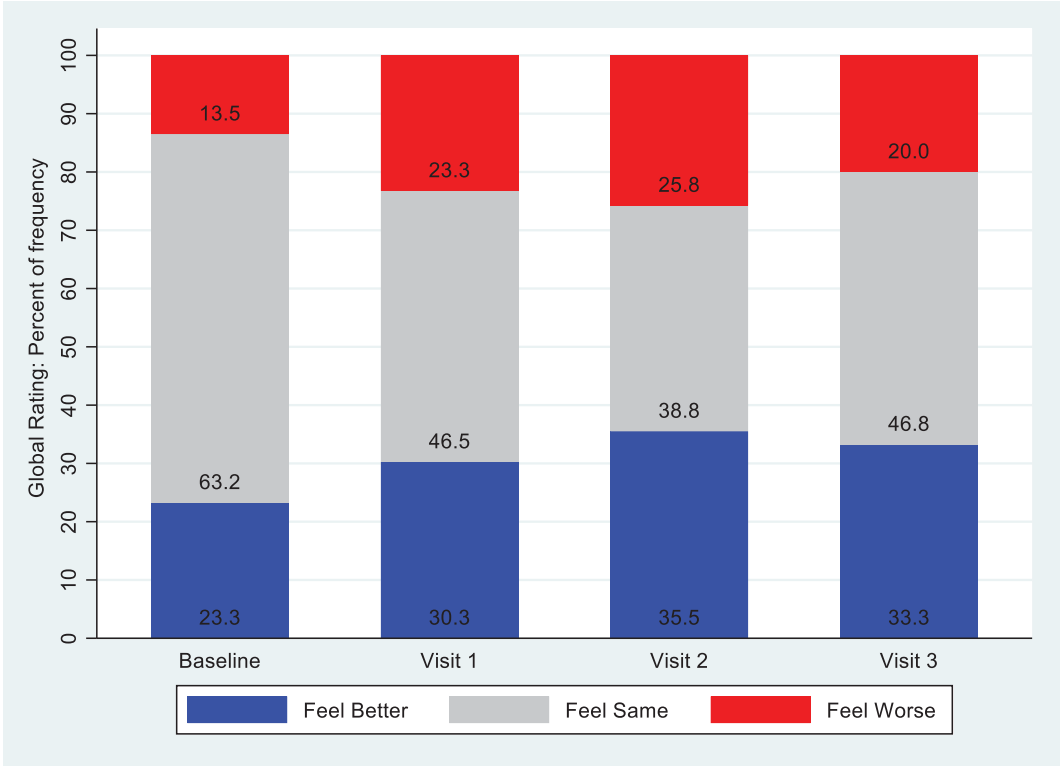
**Table A2.1: Baseline characteristics**

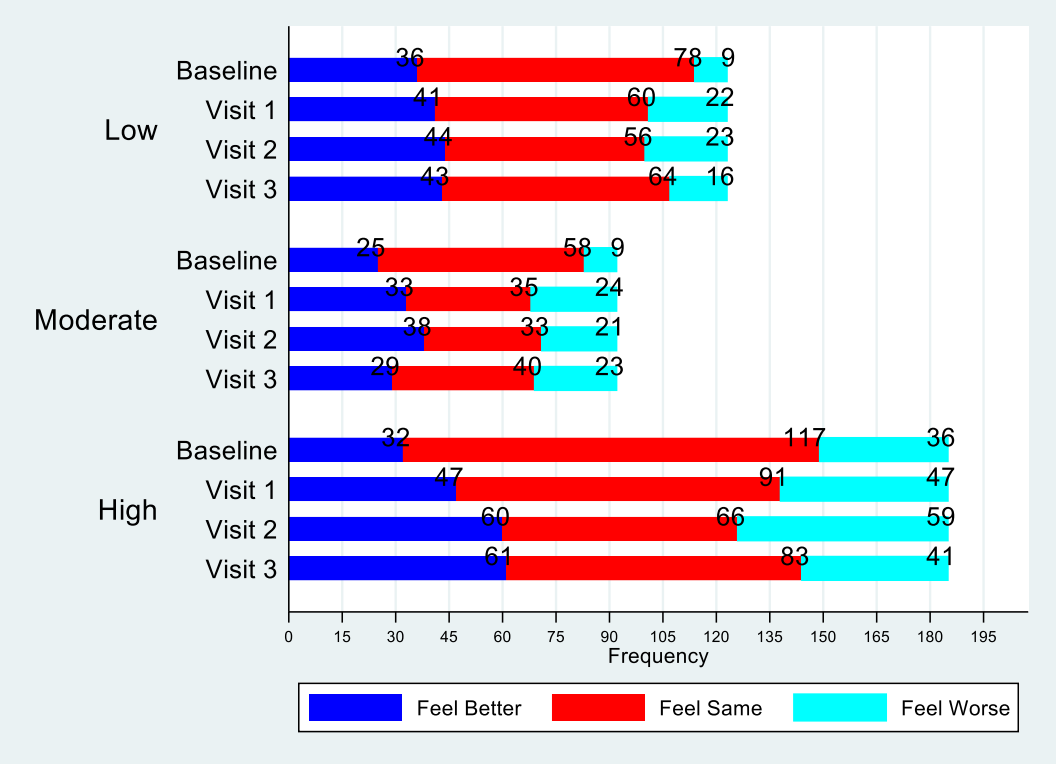
| Characteristic | N (%) unless stated otherwise |
|---|---|
| Sex | |
| Female | 273 (68.3) |
| Male | 127 (31.8) |
| Ethnic group | |
| White | 391 (97.8) |
| Ethnic minority | 9 (2.3) |
| Highest qualification | |
| Non-compulsory (A Level or above) | 254 (64) |
| Compulsory or below (GCSE equivalent or no qualifications) | 146 (36) |
| Financial status | |
| Comfortable | 92 (23) |
| Doing alright | 130 (32) |
| Just about getting by | 109 (27) |
| Finding it difficult | 69 (17) |
| Marital status | |
| Married/cohabiting | 212 (53) |
| Single | 101 (25) |
| Separated/divorced/widowed | 87 (22) |
| Currently taking antidepressants | |
| Yes | 288 (72) |
| No | 111 (28) |
| Taken antidepressants in the past | |
| Yes | 350 (88) |
| No | 50 (12) |
| Age, Mean (SD) | 48.7 (12.5) |
| BDI-II score, Mean (SD) | 19.7 (11.8) |
| PHQ-9 score, Mean (SD) | 10.1 (6.7) |
| CIS-R score, Mean (SD) | 9.3 (5.6) |

**Table A2.2: Distribution of responses to the Global Rating of change scale over time**

| Global | Rating | | Occasion Baseline | Visit 1 | Visit 2 | Visit 3 |
|---|---|---|---|---|---|---|
| Feeling | Better | | 93 | 121 | 142 | 133 |
| | | (%) | (23.25) | (30.25) | (35.5) | (33.25) |
| | Same | | 253 | 186 | 155 | 187 |
| | | (%) | (63.25) | (46.5) | (38.75) | (46.75) |
| | Worse | | 54 | 93 | 103 | 80 |
| | | (%) | (13.5) | (23.25) | (25.75) | (20) |
| | **Total** | | 400 | 400 | 400 | 400 |

Figure A2.1
Distribution of response to the Global Rating of Change Scale over time






The proportion of people reporting feeling the same reduced over time from 63.3% at baseline to 46.8% at the third visit (Table A2.2, Appendix 2). These reductions were due to increases in the proportion of those who reported feeling either better or worse with the most dramatic changes occurring at the first visit. The proportion reporting feeling better increased from 23.3% at baseline to 30.3% during the first visit and remained at similar or slightly higher levels for the remainder of follow-up. The proportion reporting feeling worse also increased from 13.3% at baseline to 23.3% during the first visit and also remained at similar or slightly higher levels for the remainder of follow-up.

At baseline 46.3% (n=185) had a CIS-R score of 20 or higher and 31% (n=123) had CIS-R levels below 12 points. Among those with a CIS-R score of 20 or higher, the majority (63%, n=117) reported feeling the same and 17% (n=32) reported feeling better compared with the two weeks previously. Among those with moderate (n=78) and low CIS-R score (n=58) a majority of 63% reported feeling the same Among those with low CIS-R score 29% (n=36) reported feeling better. Among those with moderate CIS-R 27% (n=25) reported feeling the same (Figure A2.1, Appendix 2).

*Figure SA1a: Distribution of change in PHQ9 scores for those reporting "Feeling better" and those reporting "Feeling the same" according to baseline CIS-R strata with the MCID depicted.*

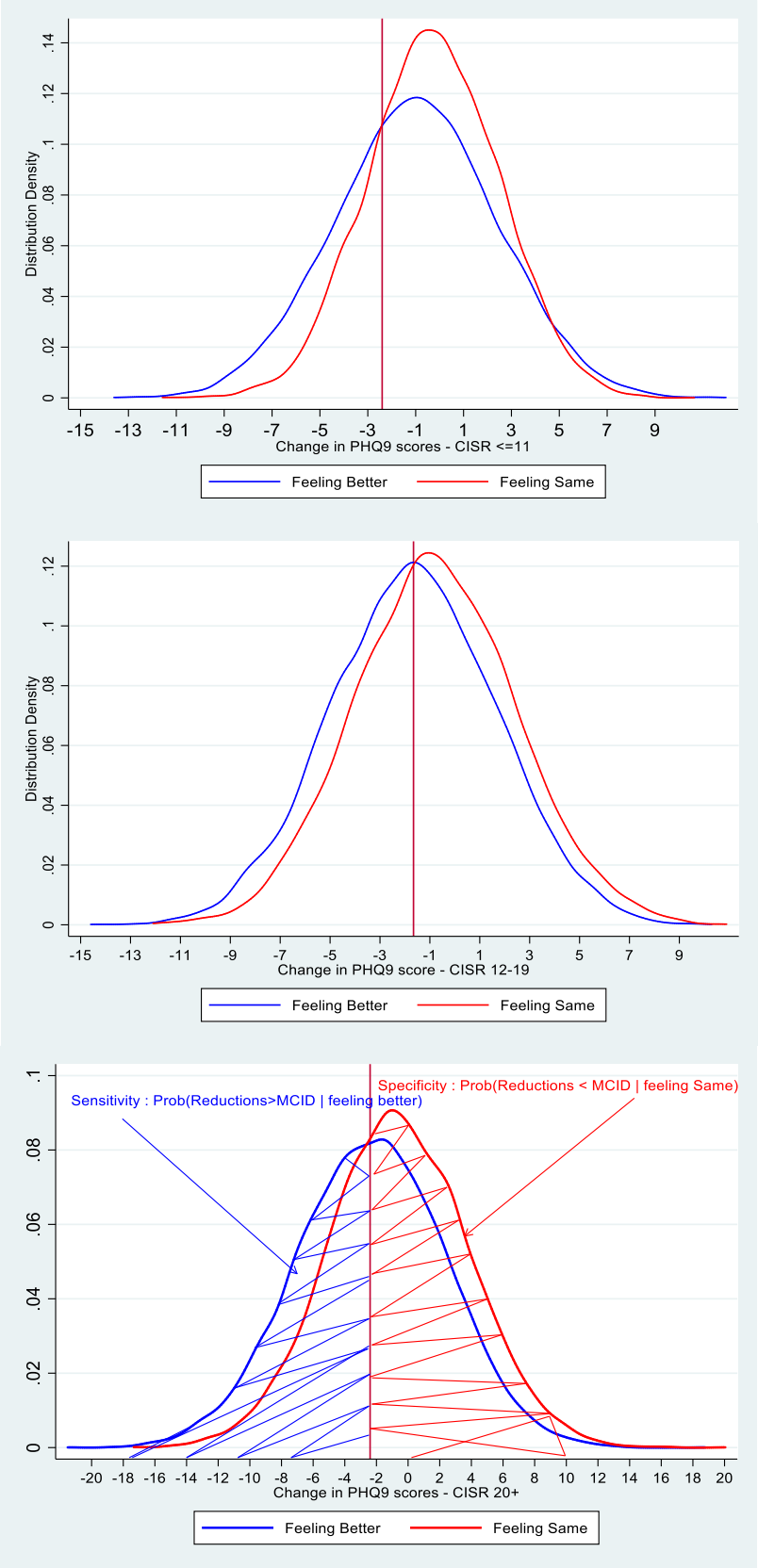


*Figure SA1b: Distribution of change in BDI-II scores for those reporting "Feeling better" and those reporting "Feeling the same" according to baseline CIS-R strata with the MCID depicted.*

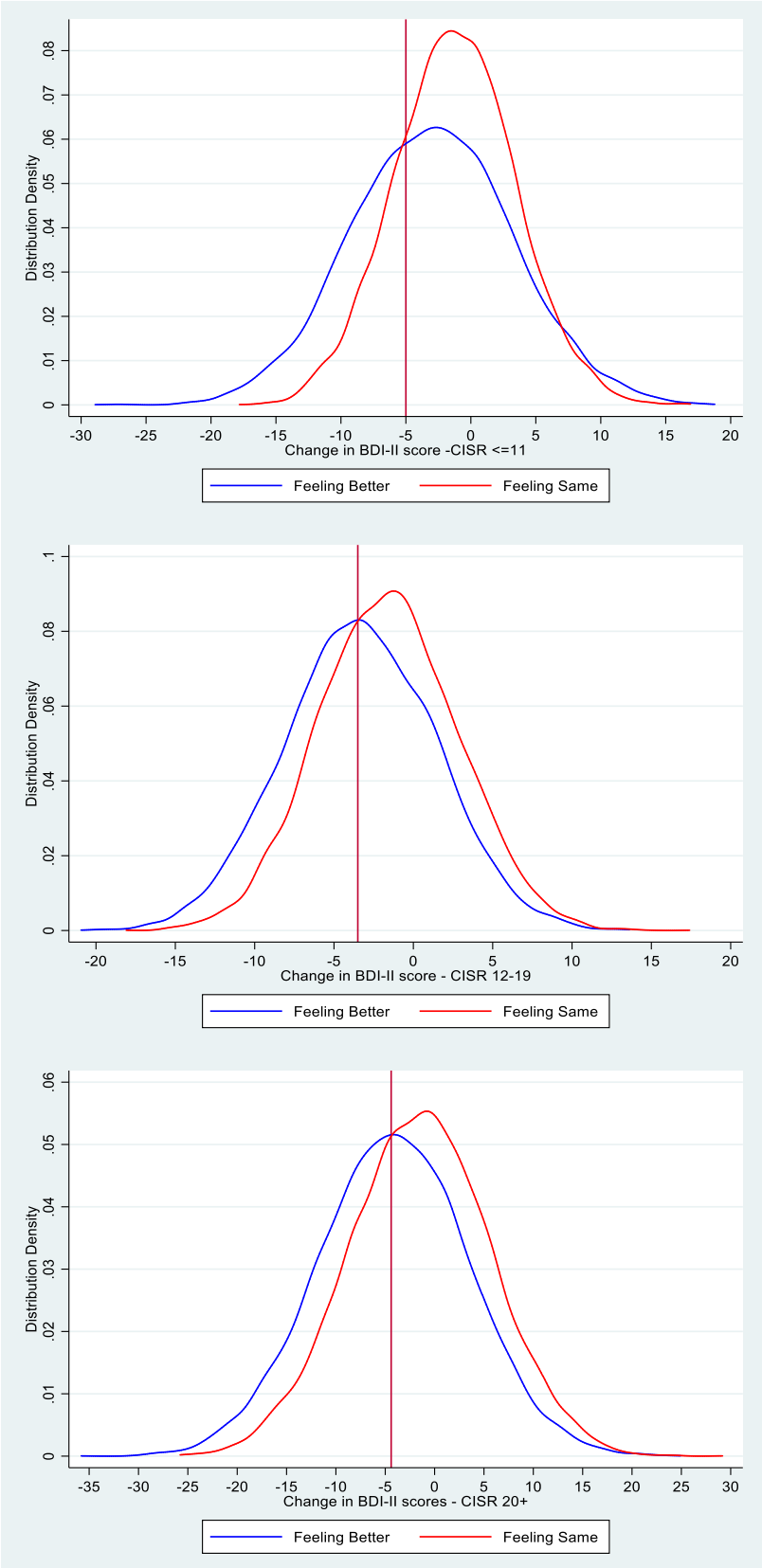
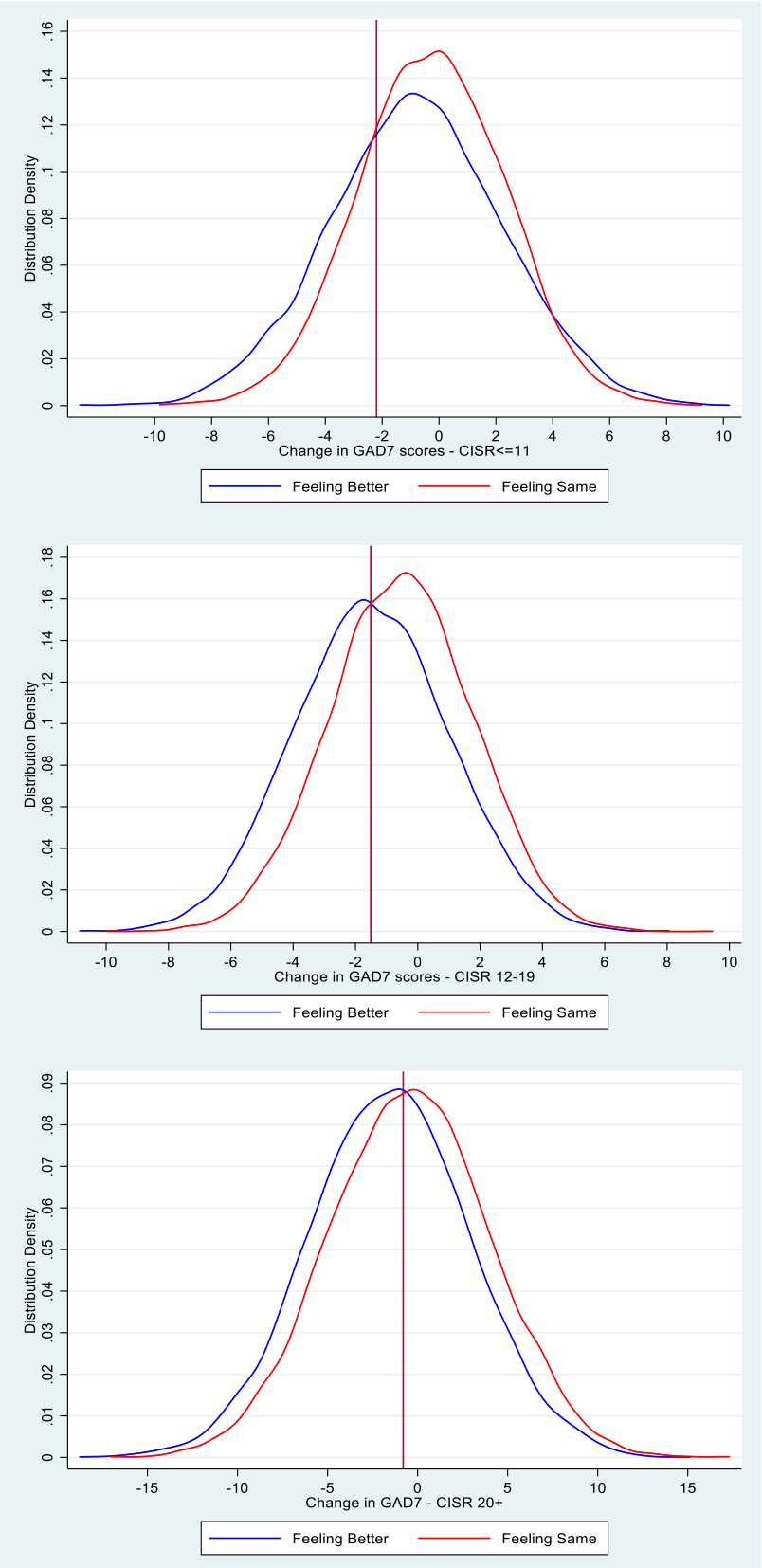


*Figure SA1c: Distribution of change in GAD-7 scores for those reporting "Feeling better" and those reporting "Feeling the same" according to baseline CIS-R strata with the MCID depicted.*



# Appendix 5 Usefulness of the PHQ-9 in primary care to determine meaningful symptoms of low mood: a qualitative study

See Malpass *et al.*[39]

# Appendix 6 Variation in recognition of happy and sad facial expressions and self-reported depressive symptom severity: a prospective cohort study

See Bone *et al.*[40]

# Appendix 7 Variation in the recall of socially rewarding information and depressive symptom severity: a prospective cohort study

See Lewis *et al.*[41]

# Appendix 8 Why are there discrepancies between depressed patients' Global Rating of Change and scores on the PHQ depression module? A qualitative study of primary care in England

See Robinson *et al.*[42]

# Appendix 9 Comparison between self-administered depression questionnaires and patients' own views of changes in their mood: a prospective cohort study in primary care

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).

# Appendix 10 A randomised controlled trial assessing the severity and duration of depressive symptoms associated with a clinical significant response to sertraline versus placebo, in people presenting to primary care with depression (PANDA trial): study protocol for a randomised controlled trial

See Salaminios *et al.*[43]

# Appendix 11 The clinical effectiveness of sertraline in primary care and the role of depression severity and duration (PANDA): a pragmatic, double-blind, placebo-controlled randomised trial

See Lewis *et al.*[44]

# Appendix 12 Cost-effectiveness of sertraline in primary care according to initial severity and duration of depressive symptoms: findings from the PANDA randomised controlled trial

See Hollingworth *et al.*[45]

EME
HS&DR
HTA
**PGfAR**
PHR

Part of the NIHR Journals Library
www.journalslibrary.nihr.ac.uk



***Published by the NIHR Journals Library***