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]

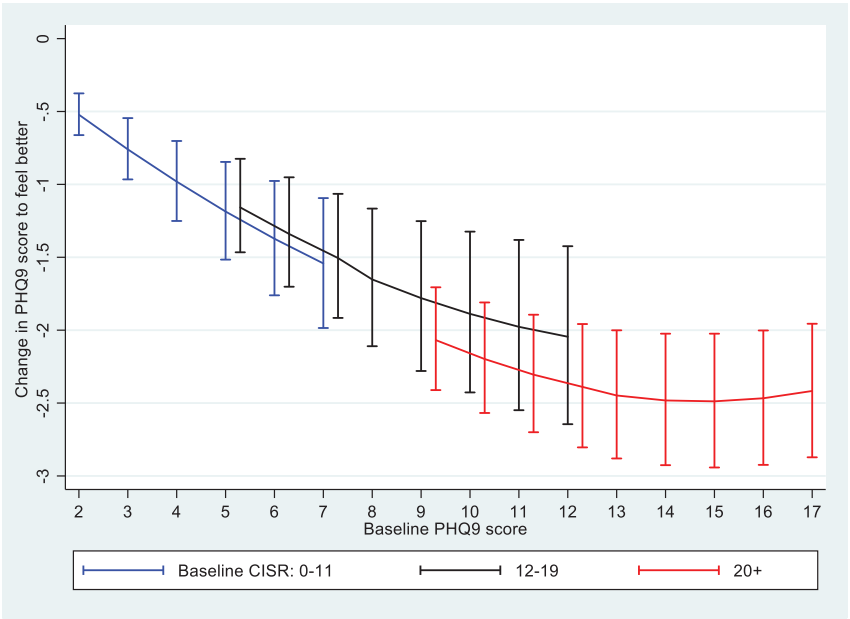
[(1)] SD Feeling Same/SD Feeling Better

[(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.



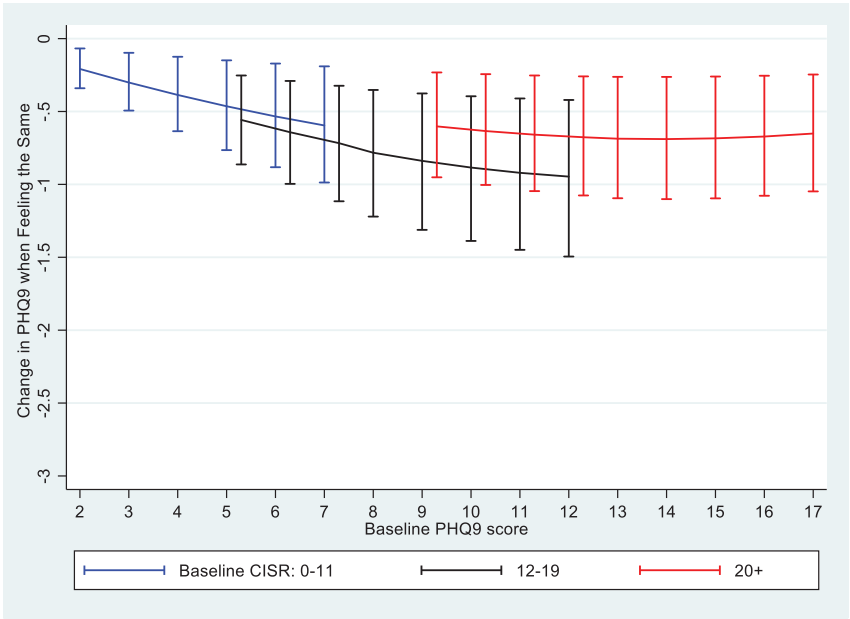


Figure 1b: Plots of estimated average absolute change in scores of those reporting feeling better (left panel) and those reporting feeling the same (right panel) for BDI-II according to their baseline scores controlling for baseline symptom severity assessed by the CIS-R.

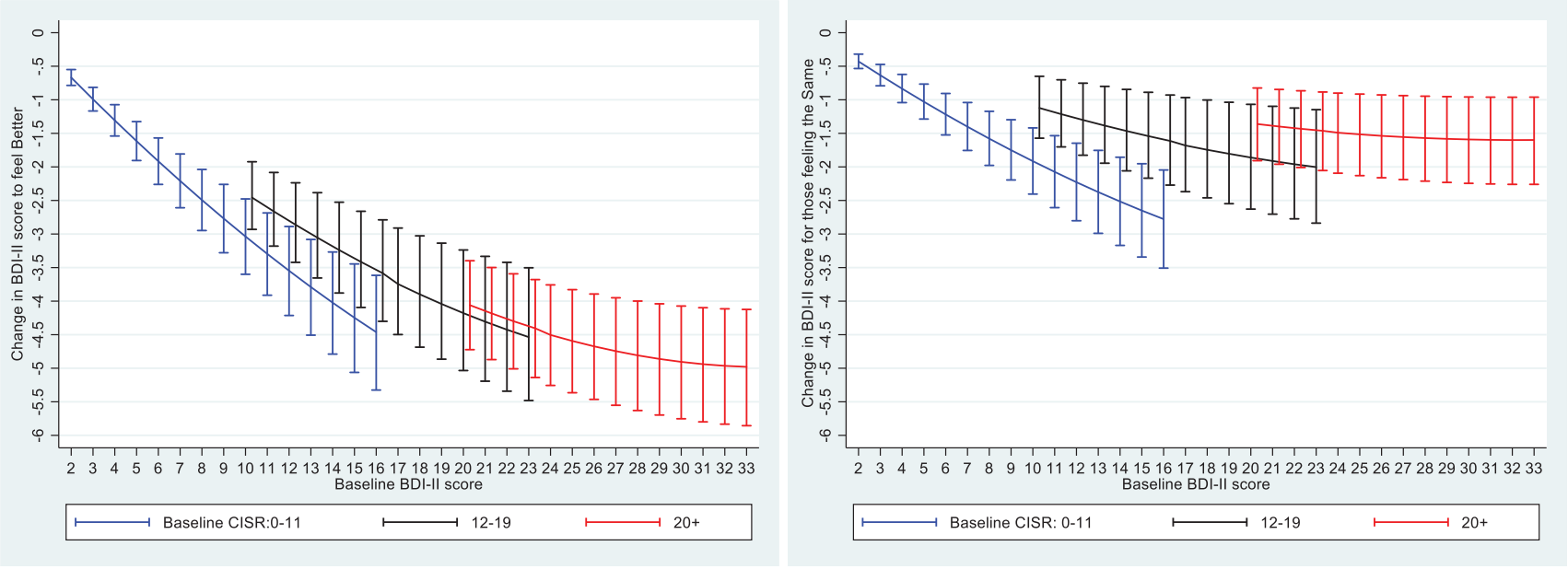
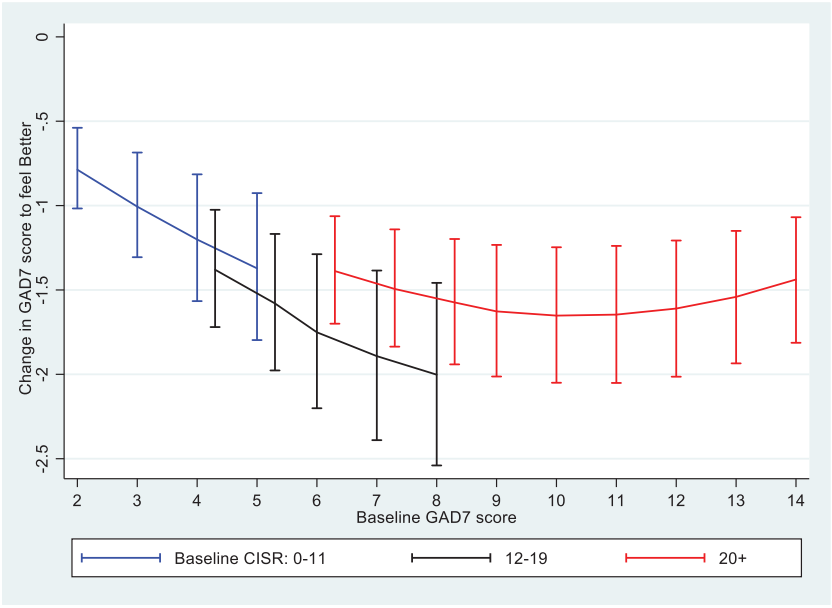
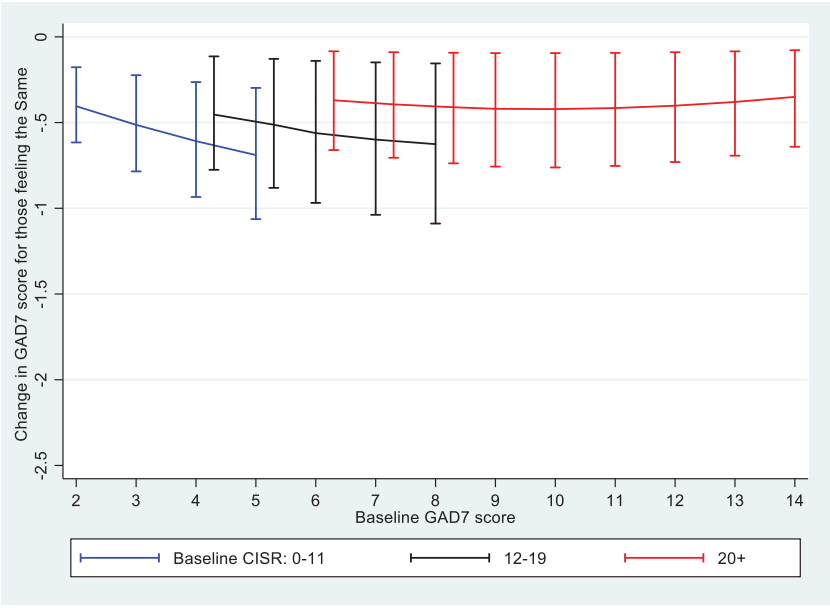


Figure 1b: Plots of estimated average absolute change in scores of those reporting feeling better (left panel) and those reporting feeling the same (right panel) for GAD-7 (panel c) according to their baseline scores controlling for baseline symptom severity assessed by the CIS-R.





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 μ_{ij} is the conditional mean and ϕ_{ij} can be interpreted as a precision parameters, in the sense that for fixed values of the mean μ_{ij} larger value of ϕ_{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.

σ_1^2, σ_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

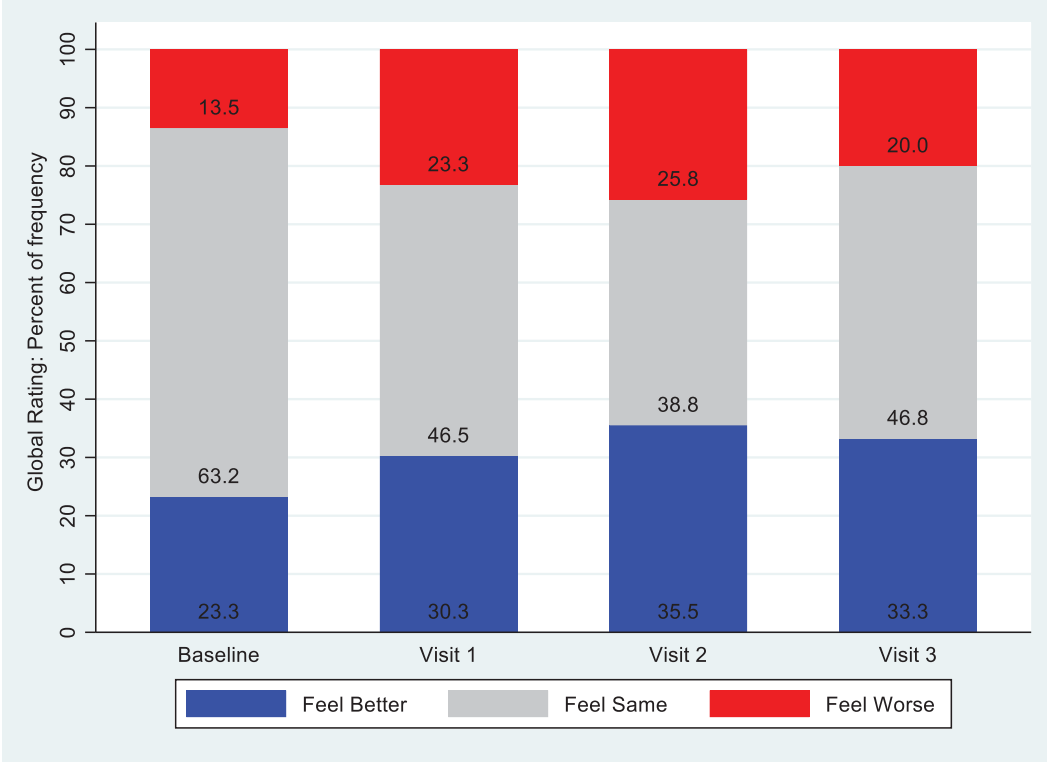
Table A2.1: Baseline characteristics	
Characteristic	**N (%) unless stated otherwise**
Sex	
Female	273 (68.3)
Male	127 (31.8)
Ethnic group	
White	391 (97.8)
Ethnic minority	9 (2.3)
Highest qualification	
Non-compulsory (A Level or above)	254 (64)
Compulsory or below (GCSE equivalent or no qualifications)	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)

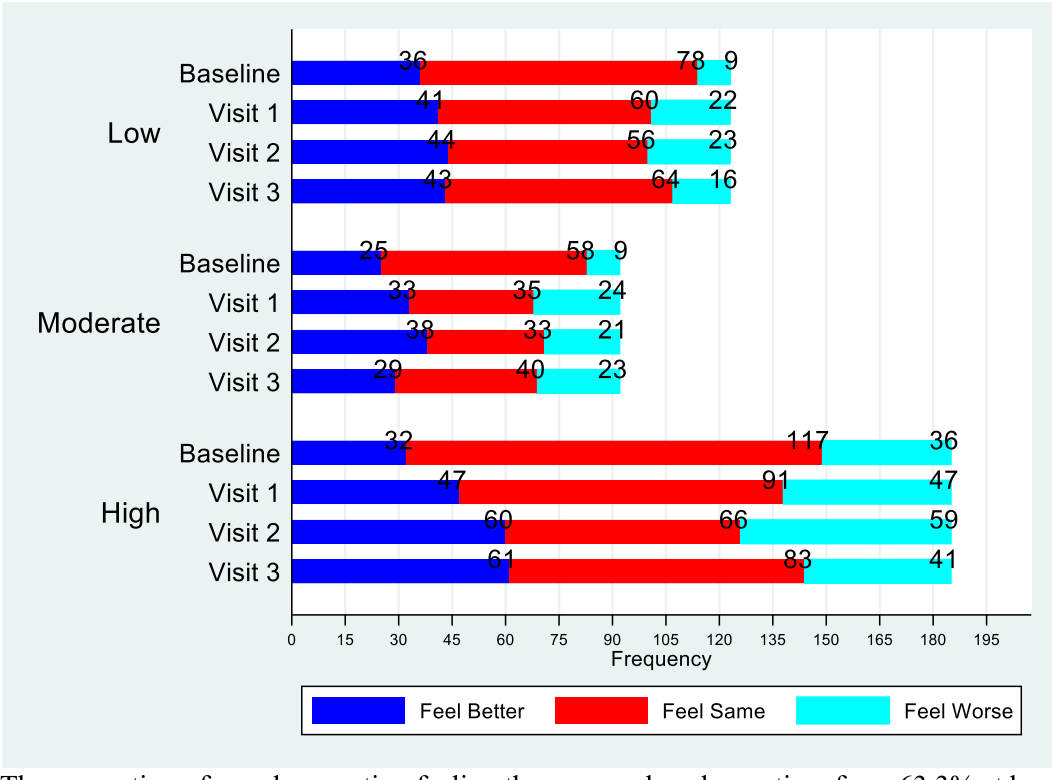
Table A2.2: Distribution of responses to the Global Rating of change scale over time

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.

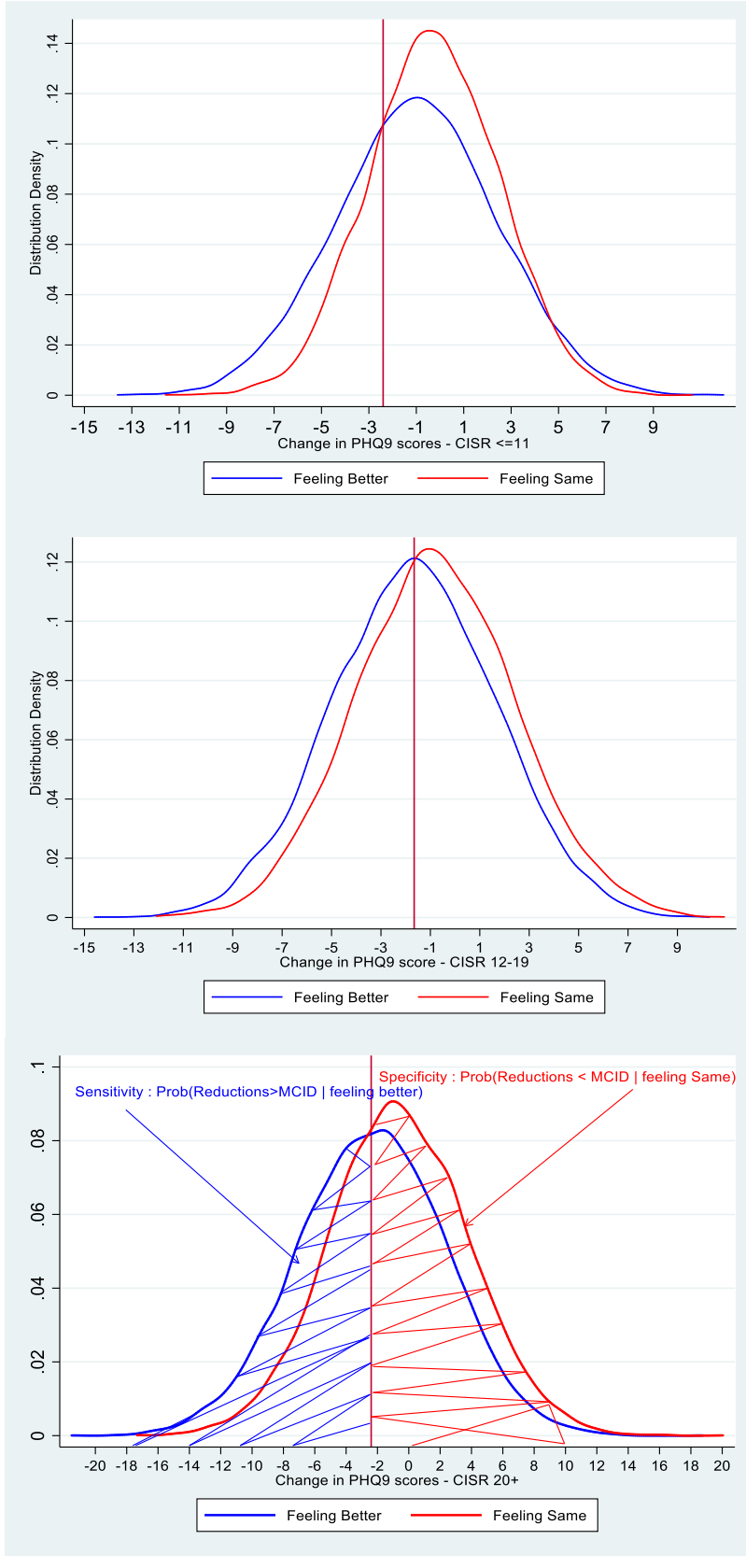


Figure SA1b: Distribution of change in BDI-II scores for those reporting "Feeling better" and those reporting "Feeling the same" according to baseline CIS-R strata with the MCID depicted.

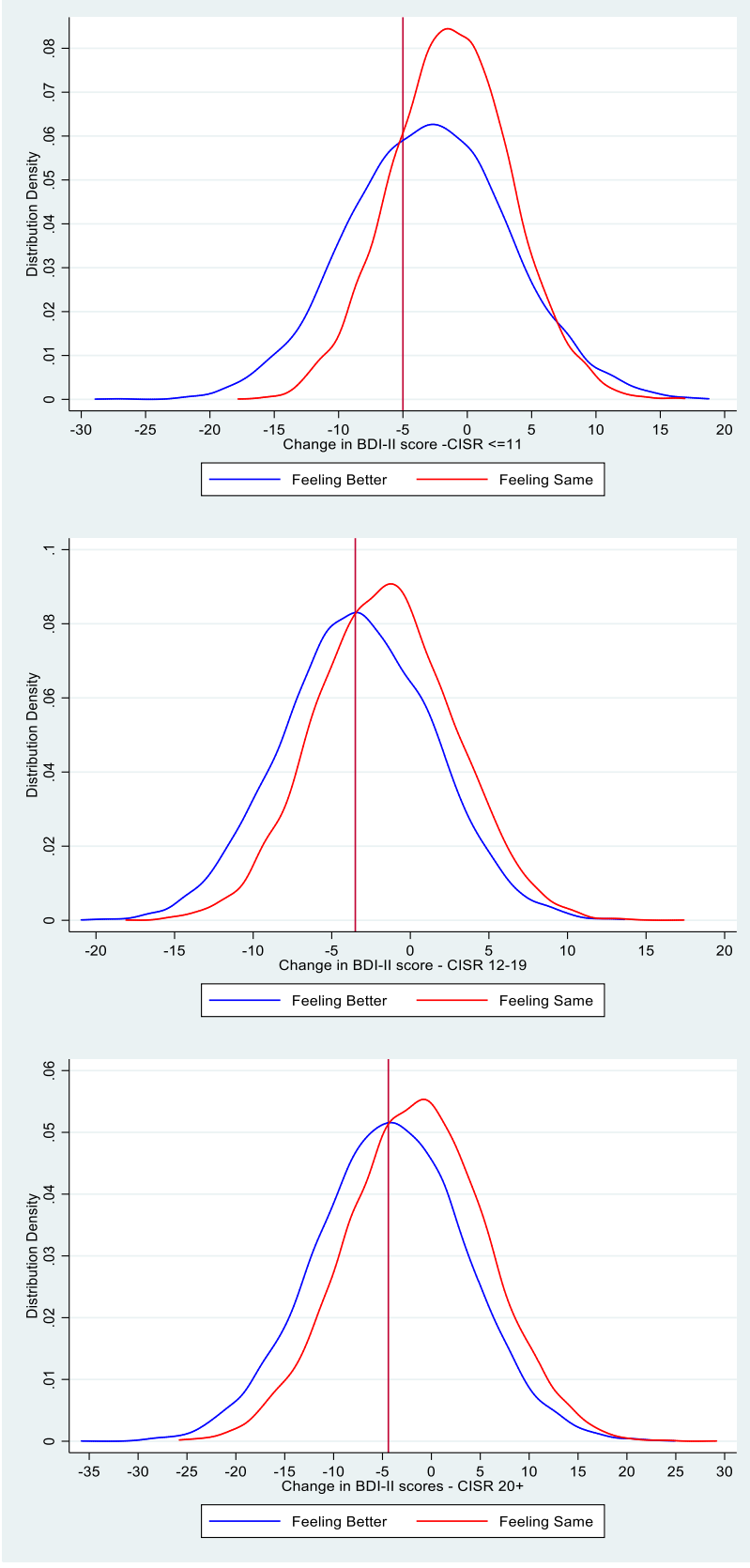


Figure SA1c: Distribution of change in GAD-7 scores for those reporting "Feeling better" and those reporting "Feeling the same" according to baseline CIS-R strata with the MCID depicted.