Concordance between PHQ-9 scores and patients’ experiences of depression: a mixed methods study

Alice Malpass, Alison Shaw, David Kessler and Deborah Sharp

ABSTRACT

Background
In 2009, a new indicator (DEP 3) was introduced into the Quality and Outcomes Framework. GPs are now encouraged to assess response to antidepressant treatment 5–12 weeks after the initial assessment, to guide clinical decision making. The Patient Health Questionnaire (PHQ-9) is one of the validated instruments that GPs can use to assess the patient’s clinical state.

Aims
To explore the extent to which changes in PHQ-9 score over time reflect patients’ accounts of their experiences of depression during the same period; and to explore patients’ experiences of using the PHQ-9 within primary care consultations.

Design of study
Mixed methods.

Setting
Primary care.

Method
Patients were recruited through six GP practices. The PHQ-9 and in-depth interviews were used at the same three time points over a 6-month period during a new or first episode of depression.

Results
Patterns in the total PHQ-9 score broadly reflected patients’ accounts of the severity of their depression over time. However, the PHQ-9 was inaccurate in its assessment of the presence and intensity of thoughts of self-harm, and missed symptoms that are meaningful to patients. At the diagnostic primary care consultation, patients viewed their score as a ‘tangible’ measure of their condition. Some patients requested the PHQ-9 subsequently as a way to measure their own treatment response and recovery process.

Conclusion
The potential therapeutic value of the PHQ-9 may be dependent upon the GP’s willingness to openly discuss the results and what they may mean for the patient.

Keywords
decision making; depression; qualitative research; severity of illness index.

INTRODUCTION

Screening questionnaires have been widely advocated to aid the detection of cases of depression and thus clinical decision making, as part of an integrated treatment approach. In the UK, the National Institute for Health and Clinical Excellence (NICE) recommends standardised instruments to improve the recognition and management of depression. In the UK, the Quality and Outcomes Framework (QOF) for primary care rewards the routine detection and assessment of depression. The Patient Health Questionnaire (PHQ-9) is a validated tool assessing severity and treatment response. It comprises nine questions based on the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) diagnostic criteria. The PHQ-9 has been validated in different settings with various patient groups and compared to other self-rating scales.

In 2009, a new indicator (DEP 3) was introduced into the QOF. GPs are now encouraged to assess response to antidepressant treatment 5–12 weeks after the initial assessment, to guide clinical decision making. However, the authors are aware of only one study that considers the sensitivity to change over...
time of the PHQ-9. This study found that changes in
the PHQ-9 score correspond with changes in
depression diagnostic status over time.13 While this
preliminary evidence is encouraging, no qualitative
work has yet examined whether changes in PHQ-9
scores reflect patients’ reported experiences of their
low mood. Patients’ insights into their illness
experiences can ‘provide valuable information to
promote a concordant approach in decision
making’.14 It is therefore important to ask whether the
PHQ-9 measures changes in the severity of
symptoms during an episode of depression that are
meaningful for patients.

The aim of this study was to explore the extent to
which changes in PHQ-9 score over time reflect
patients’ accounts of their experiences of depression
during the period of the study. A secondary aim was
to explore patients’ experiences of using the PHQ-9
within primary care consultations.

METHOD

Study design
In this longitudinal, mixed methods study, the PHQ-9
and in-depth interviews were used to gather data
from primary care patients at the same three time
points over a 6-month period, during a new or first
episode of moderate to severe depression. Patients
were aged 18–75 years and had a baseline PHQ-9
score of >10.

Recruitment and sampling
Patients were recruited from six general practices in
Bristol using two methods: GP referral at the end of
a consultation in which antidepressants were
prescribed, and data searches for adults consulting
for a new episode of depression. Patients expressing
interest in the study were telephoned by the
researcher to obtain initial verbal consent, with
formal written consent obtained face-to-face prior to
the start of the first interview. Patients identified by
the GP as severely mentally ill, unable to participate
in face-to-face interviews in their own home, or in an
interview in English, were excluded.

As recruitment was slow, plans to sample on the
basis of sociodemographic characteristics and
symptom severity were not feasible. Consequently, a
pragmatic approach was adopted while seeking to
include some heterogeneity among the patients
recruited, for example regarding sex and age.
Starting or continuing antidepressant treatment was
not a prerequisite for participation, as this study was
interested in patients with a range of experiences
and preferences.

Data collection
In-depth interviews were conducted in the patient’s

Box 1. Topic guide for interviews.

- Sociodemographic background
- General health history
- Reasons for consulting the GP
- Expectations of the consultation
- Prior use of self-help
- Feelings about a diagnosis of depression
- Experience of depression and its treatment
- Experience of the GP consultation, including use of the PHQ-9
- Expectations, views, and experiences of antidepressants
- Understanding and views of plans for ongoing care

home as soon after the initial diagnostic consultation
as possible, and again at 3 and 6 months. These time
points were chosen because patients who benefit
from taking antidepressants will have experienced
some improvement in symptoms by 3 months, and
after 6 months there is likely to be a review of their
need for continued treatment. The PHQ-9 was
completed by participants in the diagnostic
consultation with their GP and at the beginning of
each research interview. All interviews were
audiotaped and transcribed verbatim.

Flexible topic guides were used in the interviews
(Box 1). As well as exploring the patient’s experience
of their depression in their own words, the study
focused some questions around four broad concepts
that would be meaningful for a comparison with
PHQ-9 data: function; engagement; mood; and
cognition (Box 2).

Data analysis
Analysis involved four stages, assisted by the
software package ATLAS.ti. (version 5.2.6). First,
drawing on the principles of constant comparison,15
the three sets of interview data from each patient

How this fits in

GPs are now encouraged to assess response to antidepressant treatment
5–12 weeks after the initial assessment, to guide clinical decision making, but
little is known about whether PHQ-9 scores reflect patients’ reported
experiences of their low mood. This study found that changes in global PHQ-9
score broadly correspond to patients’ accounts of changes in experiences of
symptom severity during an episode of depression. The PHQ-9 can be useful in
validating patients’ illness experiences and assessing their own treatment
response, particularly if discussed with their GP. However, it may be inaccurate
in its assessment of the presence and intensity of thoughts of self-harm and
may miss symptoms that are meaningful to patients.
were analysed together to look at patients’ accounts of their experiences over time, including changes in descriptions of the severity of depression. Second, a modified framework approach was adopted, and a Microsoft ‘Excel’ chart developed for each participant, with column headings labelled as engagement, mood, function, and cognition (the four concepts that map onto the PHQ-9 domains), and row headings corresponding to the three interview time points. Drawing on the coding process from stage one, the emerging themes were organised into the appropriate cells of the grid in terms of the four concepts (engagement, mood, function, cognition) at each time point. Third, the three sets of PHQ-9 scores for each patient (gathered at each interview), as well as changes in scores for each of the nine items within the PHQ-9, were studied for patterns over time. Finally, drawing on a conceptual framework for mixed methods studies, the concept of ‘complementarity’ was used to guide comparisons of patterns in the PHQ-9 scores with patients’ interview accounts of changes in severity at the same three time points (using the charts from stage 2). Complementarity refers to the use of results of one method to elaborate the results of the other, as distinct from the concept of triangulation, whose purpose is to ensure direct corroboration between data sources.

In stage one, two researchers both independently read a subset of interview transcripts to develop emerging thematic codes, which were then agreed as a team. In stage four, a subset of PHQ-9 and interview data charts were examined, to provide a clinical perspective on the extent to which patients’ accounts of severity converged with patterns in PHQ-9 scores.

RESULTS
Nine GPs and ten patients from the six participating practices were recruited between March and December 2008. All referred patients agreed to participate in research interviews. For a profile of patients, see Table 1. While the sample was small, it was felt that obtaining longitudinal data for each participant would allow the study aims to be addressed.

The results section presents patients’ accounts of the severity of their depression (in terms of the four areas of engagement, mood, function, and cognition), compared with their PHQ-9 scores of severity, highlighting areas of convergence and mismatch. It then reports patients’ views of using the PHQ-9 within the GP consultation.

Patients’ accounts of their depression severity over time
If patterns in the global PHQ-9 scores are compared with thematic patterns in the in-depth interview data at the three time points, there is convergence between the two. Changes in the experience of symptom severity reflect the shifts in overall PHQ-9 score over time in all patients. This remains the case regardless of whether the PHQ-9 score stays constant over time, reduces, or increases in severity. Box 3 illustrates the nature of this convergence, with a case study of one patient who presented with a PHQ-9 score of 23 at her initial interview, which
reduced to 16 after 3 months and to 12 after 6 months of treatment with antidepressants. The quotations illustrate the patient’s account at the three time points.

When patterns over time in individual items within the PHQ-9 are compared to the four categories of engagement, function, mood, and cognition, more divergence occurs. Broadly speaking, patients’ accounts of their mood were consistently represented in the changes in PHQ-9 itemised scores. There was one clear account of ‘function’ not matching PHQ-9 itemised scores.

Accounts of cognition and sense of engagement matched least well with itemised scores of the PHQ-9. An example of a mismatch in descriptions of engagement occurs in a patient whose responses to item 1 on the PHQ-9 suggested that she was consistently anhedonic. This did not reflect the shift in enjoyment of social engagement over time: She had felt ‘a bit antisocial’ at the first interview, and that she ‘really can’t be bothered’ at the 3-month interview. However, by the 6-month interview she described feeling ‘keen’, ‘making a bit of an effort and looking forward’ to meeting up with people (105C).

The mismatch between patients’ descriptions of their cognition and patterns in related individual items within the PHQ-9 is more complex. Two important areas of divergence are presented.

**Thoughts of self-harm and suicide**

First, there was a mismatch between the PHQ-9 and qualitative data regarding ideas about suicide. There were instances where patients discussed suicidal ideation during interview but did not record on the PHQ-9 that they wished they were dead or had thoughts about self-harm. For example, one patient (101C) describes thoughts about jumping out of a moving car and setting her bedroom on fire during the 3-month interview, but scored zero for thoughts of self-harm on the PHQ-9 (completed at the start of the interview). More subtly, another patient (104C) describes noticing a car was driving very close to hers and ‘not caring’ if an accident happened, but did not acknowledge this thought in their response to question 9 of the PHQ-9.

**Patients’ experiences of suicidal thoughts not reflected on the PHQ-9**

Second, the PHQ-9 item for thoughts of self-harm focuses on the frequency of these thoughts rather than their intensity. This distinction was not always reflected in patients’ own experiences of feeling at risk, as described during interviews. One patient (110D) reported almost daily suicidal thoughts in the initial and 3-month interviews. By 6 months, his suicidal thoughts were less frequent — only one episode lasting a few days in the previous 3-month period. However, during his 6-month interview, he...
The fantastic thing about these tablets is that it really does take the panic away.' (103C)

Ability to relax

"Usually my teeth grit together... I can’t sleep. You find when you’re on them they do relax you.’ (104C)

Tense

"It feels like butterflies, in my stomach all the time, I feel like, quite tense.’ (105C)

Patients’ experiences of using the PHQ-9 in primary care consultations

As well as completing the PHQ-9 at the start of each
interview, patients completed the PHQ-9 in their initial GP consultation. Some patients requested that their GP repeat the PHQ-9 later during their 6-month follow-up period, as a way of assessing and monitoring their own response to treatment with antidepressants:

‘I can’t really remember how bad I felt and it’s only when I go back and do things like that [the PHQ-9] ... it’s a good reflection, it gives you an idea.’ (Female, 34 years old, new episode)

For other patients, completing the PHQ-9 during the diagnostic consultation was a ‘wake-up call’, prompting them to realise that something was ‘seriously wrong’:

‘I think that questionnaire was a little bit of a flashing light to me that I really needed to sort myself out a little bit.’ (Female, 32 years old, first episode)

For a participant with postnatal depression, completing the PHQ-9, alongside seeing her GP and being prescribed antidepressants, helped her acknowledge she was unwell, and alleviated her sense of guilt at not coping:

‘It’s made me step up a gear and sort myself out. I was sort of ignoring what was going on right up until the point where we did that questionnaire and I left and had a prescription in my hand and that was a real wake-up call for me, that actually things had got quite bad and I needed to do something about it because I just didn’t have a clue ... Only at that point did I start thinking the fact that I’ve been signed off work, I might have a valid reason, I feel like a complete fraud up until then. But I allowed myself to understand it ... I’m conscious that actually what’s happening is happening to me now, rather than it has happened to someone else and it’s kind of going on around me. And that, the trigger for that was really seeing my GP and doing that questionnaire and thinking “Christ this is actually happening to me now and I need to deal with it now”’. (Female, 40 years old, first episode)

GPs are keen for the PHQ-9 not to replace the clinical conversation, but patients benefited from using the PHQ-9 in ways that were distinct from the GP–patient dialogue:

Researcher: ‘Did the GP ask you those same direct questions?’
Participant: ‘She had asked them in a round-
about way a little bit before, but actually seeing them written down with a point score ... it was a bit scary thinking “actually I’ve let this get far enough where I’m actually really miserable now”.’ (Female, 32 years old, first episode)

However, if the GP fails to discuss the PHQ-9 results and what they may mean for the patient, this may jeopardise its usefulness and may even make the patient feel worse:

‘When I did it [the PHQ-9] in there [the consultation], it [the score] was so bad that he flicked his screen off very quickly so I couldn’t actually see the result. I could see it come up with a really large number, and it said ... I remember just glancing at some awful text that came up, and he kind of went “Oh right, okay” and just kind of switched it off very quickly before I could see it ... I guess it helped erm ... it’s quite horrible doing it ... made me feel really bad as well, when you’re filling it in and, you know, it’s like most of the time “Yes”, it just is pretty horrible ... I suppose just for the purposes of me realising perhaps that I am ... it is something tangible and you are ill, you know, you’re obviously not right if you’re answering the ... you know, if you’re giving these sort of answers, he didn’t really share the results with me as such, you know other than to ... obviously start me on antidepressants I guess erm ... and that made me feel bad, the fact that he clearly didn’t want me to see what it said made me feel quite bad as well, because I thought “God it must be really bad then”.’ (Male, 35 years old, first presentation)

DISCUSSION
Summary of main findings

When comparing PHQ-9 scores and interview accounts, patterns in the total PHQ-9 score broadly reflected patients’ accounts of the severity of their depression over time. However, the PHQ-9 missed some important symptoms of low mood that are meaningful for patients’ sense of wellbeing, such as libido, apathy, or the ability to make changes. Patients may not always record their suicidal ideation accurately on the PHQ-9, and when they do, the PHQ-9 focuses on the frequency of symptoms, which may not reflect patients’ own experiences of feeling at risk from suicidal thoughts and may be more associated with the intensity of those thoughts.

Patients experienced completing the PHQ-9 at the diagnostic primary care consultation as helpful, the score being viewed as a ‘tangible’ measure of their
condition. While the process of completing the questionnaire was not always comfortable for patients, and their diagnostic score was often experienced as ‘shocking’, this often instigated a process of acknowledgment that something was wrong and acceptance that action needed to be taken — although not necessarily starting antidepressant treatment. Some patients requested the PHQ-9 subsequently as a way to measure their own treatment response and recovery process. However, the potential therapeutic value of the PHQ-9 may be dependent upon the GP’s willingness to openly discuss the results and what they may mean for the patient.

**Strengths and limitations of the study**

As far as the authors are aware, this is the first study to use mixed methods to investigate concordance between measures of depression severity according to the PHQ-9 and patients’ lived experiences of their condition, during an episode of depression. This is important if we are to understand the extent to which the PHQ-9 is sensitive to changes in symptoms over time in ways that are meaningful to individual patients.

Recruitment was slow and the sample of patients was small, representing a select group of practices, GPs, and patients. Participating patients may have been those who felt more able to talk about their experiences, and participating GPs may have had a particular interest in shared decision making or the use of screening questionnaires. GP selection bias was openly acknowledged by some GPs in the sample. However, as half of the participating patients were ‘new’, the GPs seemed willing to risk referring unknown patients.

**Comparison with existing literature**

Other studies confirm the finding that patients experience severity questionnaires as ‘validating’ their illness experience and reducing their sense of stigma.  

The mismatch between patients’ PHQ-9 scores and their lived experiences of depression — a drop in PHQ-9 scores alongside a felt sense of increasing risk of self-harm — could be concerning if GPs are encouraged by the QOF to base ongoing treatment decisions on reductions in PHQ-9 scores.  

Evidence of patients ‘gaming’ when completing severity questionnaires, either to avoid unwanted treatment outcomes (because of stigma) or to achieve their desired outcome, is relevant to the present findings. It is possible that the patients who did not record their thoughts of self-harm on the PHQ-9 were motivated by fear of stigma and unwanted treatment outcomes.

**Implications for clinical practice and future research**

There is some evidence that GPs view the PHQ-9 as a ‘tick box’ exercise in the stepped-care pathway, as intrusive to the consultation, and with limited usefulness for shared decision making and treatment choices. The study findings suggest that patients may have different views of the value of completing the PHQ-9 within consultations. Patients in this study found the process to be a productive ‘wake-up call’, useful for informing treatment decisions and providing them with a sense of their recovery over time. The PHQ-9 may assist in promoting treatment concordance, if patients find it helpful to ‘measure’ their treatment response and changes in symptom severity using such an instrument. This welcome but unintended patient-centred consequence of encouraging follow-up assessment of depression severity could valuably be examined in future research.

**Funding body**

The study was funded by the National Institute of Health Research (RK6871).

**Ethical approval**

The study was approved by the Southmead Research Ethics Committee (07/Q2002/14).

**Competing interests**

The authors have stated that there are none.

**Acknowledgements**

We would like to thank all the participating patients, doctors, and practice staff.

**Discuss this article**

Contribute and read comments about this article on the Discussion Forum: http://www.rcgp.org.uk/bjgp-discuss

**REFERENCES**


