Keding et al. BMC Medical Research Methodology (2015) 15:56
DOI 10.1186/s12874-015-0054-6

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Validity of single item responses to short message service texts to monitor depression: an mHealth sub-study of the UK ACUDep trial

Ada Keding[1*], Jan R. Böhnke[1,2], Tim J. Croudace[3], Stewart J. Richmond[4] and Hugh MacPherson[1]

Abstract

Background: An increasing number of research designs are using text messaging (SMS) as a means of self-reported symptom and outcome monitoring in a variety of long-term health conditions, including severity ratings of depressed mood. The validity of such a single item SMS score to measure latent depression is not currently known and is vital if SMS data are to inform clinical evaluation in the future.

Methods: A sub-set of depressed participants in the UK ACUDep trial submitted a single SMS text score (R-SMS-DS) between 1 and 9 on how depressed they felt around the same time as completing the PHQ-9 depression questionnaire on paper at 3 months follow-up of the trial. Exploratory categorical data factor analysis (EFA) was used to ascertain the alignment of R-SMS-DS scores with the factor structure of the PHQ-9. Any response bias with regard to age or gender was assessed by differential item functioning (DIF) analysis.

Results: Depression scores based on the PHQ-9 and R-SMS-DS at 3 months were available for 337 participants (74 % female; mean age: 42 years, SD = 11.1), 213 of which completed the two outcomes within 6 days of each other. R-SMS-DS scores aligned with the underlying latent depression of the PHQ-9 (factor loading of 0.656) and in particular its affective rather than somatic dimension. The R-SMS-DS score was most strongly correlated with depressed mood ($r = 0.607$), feeling bad about oneself ($r = 0.588$) and anhedonia ($r = 0.573$). R-SMS-DS responses were invariant with respect to gender ($p = 0.302$). However, there was some evidence for age related response bias ($p = 0.031$), with older participants being more likely to endorse lower R-SMS-DS scores than younger ones.

Conclusions: The R-SMS-DS used in the ACUDep trial was found to be a valid measure of latent affective depression with no gender related response bias. This text message item may therefore represent a useful assessment and monitoring tool meriting evaluation in further research. For future study designs we recommend the collection of outcome data by new health technologies in combination with gold standard instruments to ensure concurrent validity.

Keywords: Validity, Text Messaging, SMS, mHealth, Depression, PHQ-9, Factor Analysis, DIF, Response Bias

* Correspondence: ada.keding@york.ac.uk
[1]Department of Health Sciences, University of York, Heslington, York YO10 5DD, UK
Full list of author information is available at the end of the article

Background

Depression is a debilitating long-term health condition that is one of the leading causes of global disease burden [1, 2], and its management presents a major challenge to health care providers worldwide. As part of an emerging trend to utilise mobile devices in health care (mHealth) [3], ubiquitous mobile technologies such as short message service (SMS or text messaging) may offer a cheap and straightforward support tool to monitor outcomes in clinical care and self-management of depression and other chronic health conditions [4]. Text messaging has already been studied in the management of diabetes [5–7], asthma [8–10], lower back pain [11–13] and irritable bowel syndrome [14] for example, as well as in the support of long-term health behaviour change interventions such as weight loss [15, 16] and smoking cessation [17]. While the importance of validating health outcomes collected by text messaging has been recognised, few of the studies using SMS technology have implemented this [18, 19].

Within mental health, research has primarily focussed on utilising text messaging for the management of bipolar disorder and schizophrenia. Feasible symptom monitoring was demonstrated when gathering weekly responses of validated questionnaires for depression and mania from bipolar patients [20] and when collecting daily outcomes on several symptom dimensions from patients suffering from schizophrenia [21]. Furthermore, when employed as a low level intervention in schizophrenia, customised daily text prompts for different illness aspects improved outcomes in those areas [22], and weekly monitoring of early warning signs by patients and relatives improved rates of relapse and hospital readmission [23].

Until recently, only a small number of studies with few participants had looked specifically at the possibility of collecting depression outcomes by text message. A single item SMS subjective distress rating (scale 0 to 10) was used for daily mood monitoring in patients with anxiety or depression in a remote Australian community during and after treatment [24], and a daily SMS mood score (scale 1 to 9) was collected as an adjunct to cognitive behavioural therapy (CBT) for outpatients from different ethnic groups in the United States [25–27]. These studies found mood data collection by SMS feasible, acceptable, and predictive of PHQ-9 [28] depression scores. This has been further confirmed in a sub-study of the UK ACUDep trial [29], which collected weekly depression scores (scale 1 to 9) by text message from over 500 depressed adult participants during the first 3 months of trial follow-up [30]. The study demonstrated good response rates (94 % of patients responded to at least one text prompt, and patients replied to an average of 12.5 (SD = 3.45) of 15 texts), the depression rating correlated well with the PHQ-9 measure of depression (Kendall's tau-b = 0.570), and SMS depression scores were sensitive to change in response to the trial treatments.

Monitoring patient depression with such a simple, single SMS text score instead of the administration of lengthy questionnaires represents an attractive mode of data collection in view of compliance rates and patient burden. This is in line with other efforts to condense the measurement of depression into one or two items for the purpose of efficient patient screening and monitoring [31–34]. The choice between long and short form assessment tools will depend on the context and purpose of the evaluation, balancing ease of data collection with the need for robust clinical diagnoses [35]. It remains unknown whether a single SMS depression score, as used in the ACUDep trial, can be considered a valid measure of depression and could consequently be recommended for use in research and evaluation in clinical practice.

The present study therefore aimed to establish the validity of the ACUDep SMS depression score (termed R-SMS-DS [30]), by employing item response theory methodologies. If scores obtained for the R-SMS-DS and the PHQ-9 both measure the same latent depression variable, then this could be confirmed by including all individual items in a factor analysis. The PHQ-9 has variously been shown to be either uni-dimensional in primary care patients [36–39], or to divide into an affective and somatic dimension in certain patient populations [40–43]. It was of interest whether R-SMS-DS scores would align with either one of these dimensions if present in the ACUDep patient sample.

Depression prevalence, symptomatology and trajectories are known to differ between men and women [44–46] as well as over the course of life [45, 47]. Although the reasons for these disparities remain debated, they may be connected to differential use of health care systems [48] and important aspects of depression treatment [49]. It is therefore important that these demographic groups do not differ in the way they use the R-SMS-DS, and score differences between individuals only reflect variations in their respective levels of depression [50]. Therefore the present study also aimed to assess any response bias for the R-SMS-DS with respect to age and gender. The absence or presence of such biases will provide evidence for the relative impact of these factors on the measurement of depression with the R-SMS-DS, before it can be considered to inform valid treatment decisions in clinical practice.

Results of this study were anticipated to inform recommendations for whether and how the increasing number of research studies using mHealth technologies for patient monitoring should incorporate these tools and their validation into their study designs.

Methods

Participants

Participants included in this study took part in the ACUDep trial [29], a three arm randomised controlled trial that evaluated the effectiveness of acupuncture or counselling compared to standard care in a population of depressed adults in the North of England. Participants were 18 years of age or older, had consulted for depression within the previous five years and had ongoing depression with a score of 20 or above on the Beck Depression Inventory (BDI-II) [51]. Those recruited into the trial were invited to take part in an optional sub-study involving the use of weekly SMS text messages to monitor their depression. 755 patients were recruited into the ACUDep trial between 2009 and 2011, and 527 of these consented to the SMS sub-study.

Design

In order to investigate the validity of the R-SMS-DS [30] as a measure of depression, this study exploited the collection of the last of 15 weekly SMS text scores and PHQ-9 depression by questionnaire around the same time at 3 months follow-up of the trial. Participants were considered as a single patient group for this purpose, irrespective of their allocated trial arm. The differences in R-SMS-DS scores between treatment groups in patients' depression trajectories are reported elsewhere [30]. We used categorical data factor analysis [52] to ascertain the factor structure of the PHQ-9 in the present patient sample and the alignment of the R-SMS-DS with that structure. Following these exploratory analyses we used differential item functioning (DIF) analysis to investigate potential response bias with respect to age or gender.

Outcome measures

The PHQ-9 [28] is a nine-item depression scale based on the DSM-IV symptom criteria for major depressive disorder [53]. It is used routinely as a screening tool in clinical practice and as a standard depression severity outcome in research. Each item is scored between 0 and

Over the last 2 weeks, how often have you been bothered by any of the following problems?
(Please circle one number per row only)

	Not at all	Several days	More than half the days	Nearly every day
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself - or that you are a failure or have let yourself or your family down	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching TV	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed. Or the opposite - being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3
9. Thoughts that you would be better off dead, or of hurting yourself in some way	0	1	2	3

Fig. 1 PHQ-9 Questionnaire Wording (Source: Kroenke et al. 2001)

3, thus PHQ-9 total scores range from 0 to 27 with higher scores indicating greater depression (see Fig. 1 for complete wording of the PHQ-9). The instrument was completed by patients at baseline and follow-up on paper questionnaires, and the total score at 3 months served as the ACUDep primary endpoint.

The weekly R-SMS-DS text message sent to patients who consented to the sub-study contained the text: 'ACUDep Trial: Over the last week how depressed have you felt on average? Please reply with a score between 1 and 9; where 1 is "not at all" and 9 is "extremely"'. Up to 15 weekly text messages were sent to participants following randomisation, the final text approximately coinciding with 3 months follow-up. Received participant texts were matched to the text they were responding to, and text content was validated to arrive at a single score for each responding patient between 1 and 9, allowing half scores if patients submitted these or two adjacent scores.

Statistical analysis

Exploratory Factor Analyses (EFAs) were conducted for three groupings of ACUDep participants at 3 months follow-up: Group 1 comprised patients with complete PHQ-9 items; Group 2 were patients with complete PHQ-9 items and a valid R-SMS-DS score; and Group 3 were patients with complete PHQ-9 items and a valid R-SMS-DS score completed within 6 days of each other. Previous research suggests that PHQ-9 scores are associated with average texted mood ratings over 1 week, but not 2 weeks [27]. Group 1 was used to inform the factor structure of the PHQ-9, whereas the alignment of R-SMS-DS scores with PHQ-9 depression was explored in Groups 2 and 3, with greater agreement expected in the temporally closer assessments in Group 3.

All EFAs were computed using FACTOR 9.2 [54], using polychoric correlations in a parallel, minimum rank factor analysis with oblique (promin) rotation. One- and two-factor solutions were implemented as suggested by previous structural analyses of the PHQ-9 [36–43]. Optimal dimensionality of the item set was established, for which parallel analysis has been shown to be highly efficient [55–57]. It determines eigenvalues for random data matrices and establishes a cut-off (above 95 % based on random data) to retain relevant factors only, i.e. those that capture more common variance between the items than expected purely by chance. Item correlations between all item pairs were extracted from the analyses as well as factor loadings for the one- and two-factor models, suppressing any loadings less than 0.400. Emphasis of these analyses was on the fit of the R-SMS-DS score with the PHQ-9 factor structure.

Differential Item Functioning (DIF) with respect to age and gender was investigated by ordinal logistic regression [58, 59] in Stata version 12 [60]. The analyses included all patients with complete PHQ-9 and R-SMS-DS data at 3 months (Group 2), predicting R-SMS-DS score (values 1 to 9) from age or gender (uniform DIF) and their interaction with the PHQ-9 (non-uniform DIF, i.e. any bias that was dependent on the level of latent depression). The regression models controlled for latent depression as measured by the total PHQ-9 score at 3 months follow-up, which was expected to be highly correlated with the R-SMS-DS score, reflecting that both assess the same underlying depression construct. Evidence for response bias would be found if age, gender or their interactions with the PHQ-9 significantly ($p < 0.05$) predicted the R-SMS-DS over and above the PHQ-9 total score, potentially rendering comparisons between them unfair [50]. The direction of any identified DIF was explored, and the DIF effect size determined by comparison of pseudo R^2 values between the analysis models and a base model including PHQ-9 total score as the only predictor. Continuous variables (age and PHQ-9) were centred for all analyses.

Ethical approval and consent

Full ethical approval for the trial was granted by York NHS Research Ethics Committee on 21st September 2009 (ref: 09/H1311/75), together with research governance approval shortly thereafter from North Yorkshire & York Primary Care Trust. All participants provided informed written consent.

Results

Data availability and baseline characteristics

Of 755 randomised ACUDep trial participants, 602 patients had complete PHQ-9 data for all items at 3 months follow-up (Group 1). Of the 527 ACUDep participants who additionally consented to take part in the SMS sub-study, 373 patients responded with a valid text message to their last follow-up SMS, which broadly coincided with the 3-month PHQ-9 follow-up time point. Of these, 337 had complete PHQ-9 data (Group 2). PHQ-9 questionnaires were completed on average 8 days from responding to the R-SMS-DS (range −8 to 75 days, completion date missing for 11 patients), and 213 patients (63 %) completed these outcomes within 6 days (Group 3). Baseline characteristics for all randomised ACUDep patients and the different patient groups included in the factor analyses are given in Table 1. Apart from fewer retired patients in Group 3, the demographic profile did not substantially differ between groups.

Factor analyses

Results of all factor analyses are presented in Table 2. The initial EFA of the PHQ-9 using all available data (Group 1, $n = 602$) confirmed the uni-dimensional structure of the scale, with the first identified factor explaining 64 % of the variance and being the only one that captured more common variance than expected by

Table 1 Baseline characteristics of different analysis populations

Characteristic	Total patients in ACUDep trial $n = 755$	Group 1 Patients with PHQ-9 score at 3 months $n = 602$	Group 2 Patients with PHQ-9 score at 3 months & R-SMS-DS at 3 months (any time) $n = 337$	Group 3 Patients with PHQ-9 score at 3 months & R-SMS-DS at 3 months (±6 days) $n = 213$
Age				
Mean (SD)	43.5 (13.37)	44.7 (13.14)	42.2 (11.13)	42.5 (11.18)
Median (min, max)	43 (18, 93)	43.5 (18, 89)	42 (18, 75)	42 (18, 75)
Gender, n (%)				
Male	201 (26.6)	159 (26.4)	86 (25.5)	50 (23.5)
Female	554 (73.4)	443 (73.6)	251 (74.5)	163 (76.5)
Employment, n (%)				
Working full-time	281 (37.2)	223 (37.0)	140 (41.5)	80 (37.6)
Working part-time	144 (19.1)	116 (19.3)	63 (18.7)	47 (22.1)
Unable to work	95 (12.6)	69 (11.5)	36 (10.7)	20 (9.4)
Looking after home	83 (11.0)	62 (10.3)	37 (11.0)	25 (11.7)
Retired	65 (8.6)	61 (10.1)	15 (4.5)	10 (4.7)
Full-time education	23 (3.0)	17 (2.8)	13 (3.9)	8 (3.8)
Other	48 (6.4)	40 (6.6)	23 (6.8)	18 (8.5)
Missing	16 (2.1)	14 (2.3)	10 (3.0)	5 (2.3)
Depression, mean (SD)				
Age at 1st major episode	25.2 (12.28)	25.6 (12.51)	23.8 (11.03)	24.9 (11.66)
Baseline BDI-II	32.5 (8.72)	32.1 (8.62)	31.8 (8.54)	31.5 (8.27)
Baseline PHQ-9	16.0 (5.29)	15.7 (5.32)	15.6 (5.48)	15.4 (5.29)

Table 2 Summary of exploratory factor analysis item factor loadings[a] of PHQ-9 and R-SMS-DS scores at 3 months follow-up

	Group 1			Group 2			Group 3		
	$N = 602$ patients with PHQ-9			$N = 337$ patients with PHQ-9 & R-SMS-DS (any time)			$N = 213$ patients with PHQ-9 & R-SMS-DS (within 6 days)		
	Loadings			Loadings			Loadings		
	One-Factor	Affective	Somatic	One-Factor	Affective	Somatic	One-Factor	Affective	Somatic
PHQ-9 Item Descriptive/Variance explained	64 %	— 73 % —		61 %	— 69 % —		61 %	— 70 % —	
1. Loss of interest (anhedonia)	.852	.632	-	.855	.540	-	.854	.448	.453
2. Depressed mood	.856	.875	-	.876	.768	-	.882	.683	-
3. Sleep disturbance	.778	-	.942	.761	-	.973	.740	-	.994
4. Fatigue	.794	-	1.078	.794	-	1.041	.779	-	.831
5. Appetite changes	.718	-	.640	.700	-	.617	.719	-	.573
6. Feeling bad about oneself	.816	.972	-	.840	1.244	-	.846	1.169	-
7. Concentration difficulties	.791	.554	-	.738	.411	-	.746	-	.634
8. Psychomotor disturbance	.735	.507	-	.690	.405	-	.700	-	.714
9. Thoughts of death or self-harm	.704	.894	-	.724	.860	-	.719	.826	-
R-SMS-DS: 'How depressed have you felt?'	n/a[b]	n/a[b]	n/a[b]	.656	.501	-	.692	.616	-
Correlation between factors		— .834 —			— .820 —			—.793 —	

[a]Loadings < 0.400 suppressed
[b]Group 1 analyses excluded the SMS score, as this was not available for all patients

Table 3 Polychoric correlations between R-SMS-DS score and PHQ-9 items

PHQ-9 items	Group 2	Group 3	Grp 2 - Grp 3
	Patients with PHQ-9 & R-SMS-DS (any time) $N = 337$	Patients with PHQ-9 & R-SMS-DS (within 6 days) $N = 213$	Patients with PHQ-9 & R-SMS-DS (outside 6 days) $N = 113$
1. Loss of interest (anhedonia)	.573	.593	.545
2. Depressed mood	.607	.619	.561
3. Sleep disturbance	.474	.458	.517
4. Fatigue	.479	.487	.481
5. Appetite changes	.472	.513	.425
6. Feeling bad about oneself	.588	.665	.454
7. Concentration difficulties	.450	.485	.414
8. Psychomotor disturbance	.421	.404	.473
9. Thoughts of death/self-harm	.436	.472	.381

chance (parallel analysis). Individual item loadings were high and ranged between 0.704 and 0.856. When forced into a two-factor solution, the PHQ-9 items divided into two highly correlated (0.834) dimensions consistent with previous findings: a factor of somatic symptoms (sleep, fatigue, appetite) and a factor of affective symptoms represented by the remaining six depression items.

When including the R-SMS-DS score in the analyses (Table 3), the PHQ-9 items that correlated most strongly for any patients with both outcomes (Group 2) were depressed mood (0.607), feeling bad about oneself (0.588) and anhedonia (0.573). Correlations for the sub-set of patients whose R-SMS-DS and PHQ-9 responses were given within 6 days (Group 3) exhibited a similar pattern and were generally higher, with the exception of sleep and psychomotor disturbance. These mainly somatic depression symptoms correlated more strongly with the R-SMS-DS score when assessments were more widely spaced in time (see Table 3).

When R-SMS-DS scores were included in the factor analyses (Table 2), the one-factor structure remained the optimal description of the data (parallel analysis; 61 % explained variance). The R-SMS-DS text score loaded moderately highly onto the underlying depression factor: 0.656 in the overall model (Group 2) and 0.692 for texts within 6 days of PHQ-9 completion (Group 3). When analysed as a two-factor solution, the R-SMS-DS score aligned with the six items of the PHQ-9 affective dimension (0.501 for Group 2 patients). The two-factor structure altered slightly when using the sample of patients who responded within 6 days (Group 3): PHQ-9 items for concentration difficulties and psychomotor disturbance now loaded predominantly onto the somatic dimension, and anhedonia loaded equally onto the affective and somatic dimension. The R-SMS-DS score still aligned with the dimension made up of the remaining core affective items (0.616), comprising depressed mood, feeling bad about oneself and having thoughts of dying or self-harm. The two dimensions remained highly correlated however (0.793), and the parallel analysis identified a one-factor solution as optimal in this sample too, explaining 61 % of the variance.

In summary, the R-SMS-DS was shown to pick up on the same underlying depression as the PHQ-9, in particular the affective dimension of depression.

Response bias

Following results of the EFAs, the specified PHQ-9 total score in the logistic DIF regressions was replaced with the affective sub-score PHQ-9_A, calculated as the sum of the PHQ-9 affective items (Items 1,2,6,7,8,9). Although according to the results of the parallel analysis a one factor solution described the responses to all items, we used the PHQ-9_A as a measure with maximum uni-dimensionality, thereby providing a more concise estimate of the characteristic being measured by the R-SMS-DS than the total score. The resulting regression coefficients were expressed as odds ratios and are presented in Table 4.

The DIF analysis for age revealed no evidence for non-uniform DIF ($p = 0.271$) but some evidence for uniform age related DIF ($p = 0.031$), change in pseudo $R^2 = 0.004$. Using predicted endorsements of each R-SMS-DS value based on the regression model, we found older

Table 4 DIF ordinal logistic regression results (Group 2, $n = 337$)

Predictor	Odds ratio	SE	95 % CI	*p*
Age DIF analysis				
Age	0.98	0.009	0.96, 1.00	.031
Age x PHQ-9_A	1.00	0.002	1.00, 1.01	.271
PHQ-9_A	1.46	0.044	1.38, 1.55	<.001
Gender DIF analysis				
Gender (being female)	1.26	0.280	0.81, 1.95	.302
Gender x PHQ-9_A	1.07	0.061	0.95, 1.19	.250
PHQ-9_A	1.39	0.073	1.25, 1.54	<.001

PHQ-9_A = Sum of affective PHQ-9 items (Items 1,2,6,7,8,9)

participants being more likely to use lower scores in their text responses (R-SMS-DS scores of 1 to 3) and less likely to use higher scores (R-SMS-DS scores of 5 to 9) compared to younger participants with the same level of affective depression ($PHQ\text{-}9_A$). The DIF analysis for gender revealed no evidence for uniform DIF ($p = 0.302$) nor non-uniform DIF ($p = 0.250$), change in pseudo $R^2 = 0.002$. Thus results of the DIF analyses suggest some evidence of age related response bias but not gender bias for the R-SMS-DS.

Discussion

The present study set out to validate a single depression rating item submitted by SMS text message (R-SMS-DS) against data of the widely validated PHQ-9 concurrently collected by post, which were available for a depressed adult sub-population of the UK ACUDep trial. R-SMS-DS scores were found to correlate well with latent depression when included in a combined single-factor solution explanatory factor analysis with the individual PHQ-9 items. The most closely associated PHQ-9 items were the two core DSM-IV criteria of depressed mood and anhedonia as well as feeling bad about oneself. The correlations closely mirrored those observed for a single-item paper based depression severity rating when correlated with DSM-IV criteria in a population of psychiatric outpatients undergoing treatment for major depression [32]. With the exception of sleep and psychomotor disturbances, item correlations were larger when patients completed the two assessments closer in time, therefore results suggest that the R-SMS-DS score did indeed measure depression as desired.

While the optimal one-factor model in this study lent further support to the uni-dimensionality of the PHQ-9, it was unsurprising to find that R-SMS-DS ratings aligned with the affective rather than somatic dimension of depression in the pre-specified two-factor analyses. This raises the possibility of complementing the R-SMS-DS with one or more physical symptom questions if monitoring of the somatic depression dimension is additionally desired. Sleep, fatigue and appetite were picked up as core somatic symptoms in line with all previous studies of a two-dimensional PHQ-9 structure. Interestingly, a model with these three symptoms alone forming the somatic dimension (found in selected previous research [40, 42, 61]) was supported in patients who had both valid PHQ-9 data and patients with valid PHQ-9 and any R-SMS-DS data; whereas the most commonly observed two-factor structure [40, 41, 43, 62] with the additional two somatic items of concentration difficulties and psychomotor disturbance was only observed in the sub-set of patients whose PHQ-9 and R-SMS-DS responses were closer in time (within 6 days). The possible loading of anhedonia on the somatic dimension for these patients had previously only been recorded in one study of spinal cord injury patients at a single long-term follow-up point [40]. Patient characteristics in terms of demographics and baseline depression did not appear to differ for patients in this group, so it may be the result of differences in other patient characteristics, such as present comorbidities affecting the rating of somatic symptoms. Alternatively the model factors may be less stable in this group as the smallest analysed sub-sample.

Consistent use of the R-SMS-DS was demonstrated across men and women. However, older patients were found to be less likely to endorse higher scores even when their degree of latent depression (as defined by the PHQ-9) was indicative of such an elevated level. This could be a result of a different understanding of the 'feeling depressed' terminology used in the text message, which has been discussed in the epidemiological literature of depression both as a shift towards a more somatically driven concept or as confounding with other somatic morbidities [63, 64]. Further reasons could be different attitudes towards communicating mental wellbeing by mobile technologies or a greater reluctance to potentially arouse cause for concern. Such age bias could affect the sensitivity of the R-SMS-DS score if used for depression screening, however it is unlikely for that to be its primary use. We envisage the R-SMS-DS as a monitoring tool for patients who have already undergone formal depression assessment. The direction of the age bias was opposite to that identified in a sample of UK primary care patients for the PHQ-9 items of low mood and anhedonia for patients aged 55 and over [65]. It remains possible that the observed bias in this study is a consequence of the relatively small total sample size or the small number of older patients in the sample. While we used age as a continuous predictor, the number of patients for whom the effect was identified based on marginal effect plots was rather low ($n = 8$ participants ≥ 65 years, 2.4 %). Moreover, the magnitude of the association between age and R-SMS-DS score (OR = 0.98) was only weak [66], and the effect size in terms of pseudo R^2 [67] was negligible. The stability of this bias remains to be confirmed in a larger patient sample including a qualitative assessment of possible reasons.

Overall, results of this study add further support to the validity of collecting depression severity outcomes by SMS, which had already been shown to be feasible and acceptable in adults with ongoing depression in primary care in the ACUDep trial [30]. To our knowledge, this is the first study aiming to validate an SMS self-report tool for depression using item-response theory methodologies, and results are strengthened by the use of a gold standard validated patient self-report depression instrument (PHQ-9) based on DSM-IV criteria for comparison. Despite the relatively small sample size of this study, patients agreeing to submit weekly text messages and who were included in the present analyses were representative of those taking

part in the ACUDep trial (Table 1), who in turn were typical of adults in the UK with ongoing depression in primary care.

However, findings cannot be extrapolated to patients who are presenting with depression for the first time or who do not consult in primary care at all. A further limitation includes the temporal difference between PHQ-9 and R-SMS-DS data completion, which had not been designed to be collected concurrently, resulting in considerable between-patient variability in the time between completing the assessments. In addition, the reference time frame differed for the two measures (PHQ-9: over the last two weeks; R-SMS-DS: average over the last week), therefore it is not certain whether patients were in the same mental state when reporting those outcomes. Indeed the positive findings of this study may only represent a conservative estimate of the level of association. However, the depression outcomes linked with one another in this study were patient reported only, and no independent assessment was carried out in order to confirm clinical validity. Moreover, only the association between R-SMS-DS and a single screening tool (PHQ-9) has been demonstrated so far, and further convergent validity needs be shown in order to establish the R-SMS-DS as a valid estimate of latent depression. Capturing the full multi-faceted nature of depression will never be possible by a single item, and this is not the aim of the R-SMS-DS monitoring tool.

For future studies we suggest to include at least one assessment that allows researchers to test the concurrent validity of their novel electronic or mHealth tools with a gold standard instrument collected at the same time, an approach that has not yet been widely adopted. The shortcomings of this study could be addressed by a more controlled, dedicated design, either as standalone work or embedded in larger investigations, with particular attention to the magnitude and context of any response bias. The successful use of tools from the framework of item response theory for the validation of SMS scores at a single time point might also be extended to investigate the longitudinal validity of the R-SMS-DS scores, which had been collected weekly over 3 months. Notwithstanding such further methodological work, we believe that findings from the present and a previous study [30] have provided sufficient evidence for the feasibility, acceptability and validity of the R-SMS-DS for monitoring depression in the ACUDep study population. Given these findings, we encourage investigators and clinicians to incorporate the R-SMS-DS as a free to use outcome measure in the study of depression management in different clinical populations. If verified against other validated depression measures and found acceptable in different clinical contexts, the R-SMS-DS could be considered for use in routine clinical practice.

Conclusions

This study has demonstrated that the self-report R-SMS-DS depression item used in the ACUDep trial was a valid measure of the affective dimension of depression in this study population. In agreement with previous findings, the R-SMS-DS may therefore represent a useful assessment and monitoring tool meriting evaluation in further research.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

HMP was the chief investigator of the ACUDep trial, conceived the study design and helped to draft the manuscript. SJR was the trial manager for ACUDep. He conceived, developed and led the SMS text messaging substudy. AK performed the statistical analysis and drafted the manuscript. JRB conceived the study methodology, oversaw the data analysis and helped to draft the manuscript. TJC conceived and advised on the study methodology. All authors contributed to the manuscript and read and approved its final version.

Acknowledgements

This is independent research commissioned by the National Institute for Health Research (NIHR) under Programme Grants for Applied Research (Grant No. RP-PG-0707-10186). The views expressed in this paper are those of the author and not necessarily those of the NHS, the NIHR or the Department of Health. We would like to thank all patients who agreed to take part in this study.

Author details

[1]Department of Health Sciences, University of York, Heslington, York YO10 5DD, UK. [2]Mental Health and Addiction Research Group, Hull York Medical School, York, UK. [3]School of Nursing and Midwifery and Social Dimensions of Health Institute, University of Dundee, Dundee, UK. [4]Sydera Research Associates, Market Weighton, York, UK.

Received: 7 April 2015 Accepted: 17 July 2015
Published online: 30 July 2015

References

1. Murray CJ, Vos T, Lozano R, Naghavi M, Flaxman AD, Michaud C, et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet. 2012;380(9859):2197–223. doi:10.1016/s0140-6736(12)61689-4.
2. Whiteford HA, Degenhardt L, Rehm J, Baxter AJ, Ferrari AJ, Erskine HE, et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. Lancet. 2013;382(9904):1575–86. doi:10.1016/s0140-6736(13)61611-6.
3. Price M, Yuen EK, Goetter EM, Herbert JD, Forman EM, Acierno R, et al. mHealth: A Mechanism to Deliver More Accessible, More Effective Mental Health Care. Clin Psychol Psychother. 2014;21(5):427–36. doi:10.1002/cpp.1855.
4. Patrick K, Griswold WG, Raab F, Intille SS. Health and the mobile phone. Am J Prev Med. 2008;35(2):177–81. doi:10.1016/j.amepre.2008.05.001.
5. Benhamou PY, Melki V, Boizel R, Perreal F, Quesada JL, Bessieres-Lacombe S, et al. One-year efficacy and safety of Web-based follow-up using cellular phone in type 1 diabetic patients under insulin pump therapy: the PumpNet study. Diabetes Metabol. 2007;33(3):220–6. doi:10.1016/j.diabet.2007.01.002.
6. Kim SI, Kim HS. Effectiveness of mobile and internet intervention in patients with obese type 2 diabetes. Int J Med Informat. 2008;77(6):399–404. doi:10.1016/j.ijmedinf.2007.07.006.
7. Newton KH, Wiltshire EJ, Elley CR. Pedometers and text messaging to increase physical activity: randomized controlled trial of adolescents with type 1 diabetes. Diabetes Care. 2009;32(5):813–5. doi:10.2337/dc08-1974.
8. Anhoj J, Moldrup C. Feasibility of collecting diary data from asthma patients through mobile phones and SMS (short message service): response rate analysis and focus group evaluation from a pilot study. J Med Internet Res. 2004;6(4), e42. doi:10.2196/jmir.6.4.e42.

9. Ryan D, Cobern W, Wheeler J, Price D, Tarassenko L. Mobile phone technology in the management of asthma. J Telemed Telecare. 2005;11 Suppl 1:43–6. doi:10.1258/1357633054461714.
10. Ostojic V, Cvoriscec B, Ostojic SB, Reznikoff D, Stipic-Markovic A, Tudjman Z. Improving asthma control through telemedicine: a study of short-message service. Telemed J e Health. 2005;11(1):28–35. doi:10.1089/tmj.2005.11.28.
11. Johansen B, Wedderkopp N. Comparison between data obtained through real-time data capture by SMS and a retrospective telephone interview. Chiropract Osteopathy. 2010;18:10. doi:10.1186/1746-1340-18-10.
12. Kent P, Kongsted A. Identifying clinical course patterns in SMS data using cluster analysis. Chiropract Manual Ther. 2012;20(1):20. doi:10.1186/2045-709x-20-20.
13. Macedo LG, Maher CG, Latimer J, McAuley JH. Feasibility of using short message service to collect pain outcomes in a low back pain clinical trial. Spine. 2012;37(13):1151–5. doi:10.1097/BRS.0b013e3182422df0.
14. Brabyn S, Adamson J, MacPherson H, Tilbrook H, Torgerson DJ. Short message service text messaging was feasible as a tool for data collection in a trial of treatment for irritable bowel syndrome. J Clin Epidemiol. 2014;67(9):993–1000. doi:10.1016/j.jclinepi.2014.05.004.
15. Haapala I, Barengo NC, Biggs S, Surakka L, Manninen P. Weight loss by mobile phone: a 1-year effectiveness study. Public Health Nutr. 2009;12(12):2382–91. doi:10.1017/s1368980009005230.
16. Patrick K, Raab F, Adams MA, Dillon L, Zabinski M, Rock CL, et al. A text message-based intervention for weight loss: randomized controlled trial. J Med Internet Res. 2009;11(1), e1. doi:10.2196/jmir.1100.
17. Rodgers A, Corbett T, Bramley D, Riddell T, Wills M, Lin RB, et al. Do u smoke after txt? Results of a randomised trial of smoking cessation using mobile phone text messaging. Tobac Contr. 2005;14(4):255–61. doi:10.1136/tc.2005.011577.
18. Whitford HM, Donnan PT, Symon AG, Kellett G, Monteith-Hodge E, Rauchhaus P, et al. Evaluating the reliability, validity, acceptability, and practicality of SMS text messaging as a tool to collect research data: results from the Feeding Your Baby project. J Am Med Informat Assoc. 2012;19(5):744–9. doi:10.1136/amiajnl-2011-000785.
19. Christie A, Dagfinrud H, Dale O, Schulz T, Hagen KB. Collection of patient-reported outcomes;–text messages on mobile phones provide valid scores and high response rates. BMC Med Res Meth. 2014;14:52. doi:10.1186/1471-2288-14-52.
20. Moore PJ, Little MA, McSharry PE, Geddes JR, Goodwin GM. Forecasting depression in bipolar disorder. IEEE Trans Biomed Eng. 2012;59(10):2801–7. doi:10.1109/tbme.2012.2210715.
21. Ainsworth J, Palmier-Claus JE, Machin M, Barrowclough C, Dunn G, Rogers A, et al. A comparison of two delivery modalities of a mobile phone-based assessment for serious mental illness: native smartphone application vs text-messaging only implementations. J Med Internet Res. 2013;15(4), e60. doi:10.2196/jmir.2328.
22. Granholm E, Ben-Zeev D, Link PC, Bradshaw KR, Holden JL. Mobile Assessment and Treatment for Schizophrenia (MATS): a pilot trial of an interactive text-messaging intervention for medication adherence, socialization, and auditory hallucinations. Schizophr Bull. 2012;38(3):414–25. doi:10.1093/schbul/sbr155.
23. Spaniel F, Vohlidka P, Kozeny J, Novak T, Hrdlicka J, Motlova L, et al. The Information Technology Aided Relapse Prevention Programme in Schizophrenia: an extension of a mirror-design follow-up. Int J Clin Pract. 2008;62(12):1943–6. doi:10.1111/j.1742-1241.2008.01903.x.
24. Dunstan DA, Tooth SM. Using technology to improve patient assessment and outcome evaluation. Rural Rem Health. 2012;12:2048.
25. Aguilera A, Munoz RF. Text messaging as an adjunct to CBT in low-income populations: a usability and feasibility pilot study. Prof Psychol Res Pract. 2011;42(6):472–8. doi:10.1037/a0025499.
26. Aguilera A, Berridge C. Qualitative feedback from a text messaging intervention for depression: benefits, drawbacks, and cultural differences. JMIR MHealth UHealth. 2014;2(4), e46. doi:10.2196/mhealth.3660.
27. Aguilera A, Schueller SM, Leykin Y. Daily mood ratings via text message as a proxy for clinic based depression assessment. J Affect Disord. 2015;175:471–4. doi:10.1016/j.jad.2015.01.033.
28. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001;16(9):606–13.
29. MacPherson H, Richmond S, Bland M, Brealey S, Gabe R, Hopton A, et al. Acupuncture and counselling for depression in primary care: a randomised controlled trial. PLoS Med. 2013;10(9):e1001518. doi:10.1371/journal.pmed.1001518.
30. Richmond SJ, Keding A, Hover M, Gabe R, Cross B, Torgerson D, et al. Feasibility, acceptability and validity of SMS text messaging for measuring change in depression during a randomised controlled trial. BMC Psychiatr. 2015;15:68. doi:10.1186/s12888-015-0456-3.
31. Whooley MA, Avins AL, Miranda J, Browner WS. Case-finding instruments for depression. Two questions are as good as many. J Gen Intern Med. 1997;12(7):439–45.
32. Zimmerman M, Ruggero CJ, Chelminski I, Young D, Posternak MA, Friedman M, et al. Developing brief scales for use in clinical practice: the reliability and validity of single-item self-report measures of depression symptom severity, psychosocial impairment due to depression, and quality of life. J Clin Psychiatr. 2006;67(10):1536–41.
33. Bech P. Depressed mood as a core symptom of depression. Mediographia. 2008;30(1):9–11.
34. Löwe B, Kroenke K, Grafe K. Detecting and monitoring depression with a two-item questionnaire (PHQ-2). J Psychosom Res. 2005;58(2):163–71. doi:10.1016/j.jpsychores.2004.09.006.
35. Böhnke JR, Lutz W. Using item and test information to optimize targeted assessments of psychological distress. Assessment. 2014;21(6):679–93. doi:10.1177/1073191114529152.
36. Huang FY, Chung H, Kroenke K, Delucchi KL, Spitzer RL. Using the Patient Health Questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. J Gen Intern Med. 2006;21(6):547–52. doi:10.1111/j.1525-1497.2006.00409.x.
37. 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(546):32–6. doi:10.3399/bjgp08X263794.
38. Yu X, Tam WW, Wong PT, Lam TH, Stewart SM. The Patient Health Questionnaire-9 for measuring depressive symptoms among the general population in Hong Kong. Compr Psychiatr. 2012;53(1):95–102. doi:10.1016/j.comppsych.2010.11.002.
39. Böhnke JR, Lutz W, Delgadillo J. Negative affectivity as a transdiagnostic factor in patients with common mental disorders. J Affect Disord. 2014;166:270–8. doi:10.1016/j.jad.2014.05.023.
40. Krause JS, Reed KS, McArdle JJ. Factor structure and predictive validity of somatic and nonsomatic symptoms from the patient health questionnaire-9: a longitudinal study after spinal cord injury. Arch Phys Med Rehabil. 2010;91(8):1218–24. doi:10.1016/j.apmr.2010.04.015.
41. Elhai JD, Contractor AA, Tamburrino M, Fine TH, Prescott MR, Shirley E, et al. The factor structure of major depression symptoms: a test of four competing models using the Patient Health Questionnaire-9. Psychiatr Res. 2012;199(3):169–73. doi:10.1016/j.psychres.2012.05.018.
42. Chilcot J, Rayner L, Lee W, Price A, Goodwin L, Monroe B, et al. The factor structure of the PHQ-9 in palliative care. J Psychosom Res. 2013;75(1):60–4. doi:10.1016/j.jpsychores.2012.12.012.
43. Petersen JJ, Paulitsch MA, Hartig J, Mergenthal K, Gerlach FM, Gensichen J. Factor structure and measurement invariance of the Patient Health Questionnaire-9 for female and male primary care patients with major depression in Germany. J Affect Disord. 2015;170:138–42. doi:10.1016/j.jad.2014.08.053.
44. Piccinelli M, Wilkinson G. Gender differences in depression. Critical review. Br J Psychiatr. 2000;177:486–92.
45. Whiteford HA, Ferrari AJ, Degenhardt L, Feigin V, Vos T. The global burden of mental, neurological and substance use disorders: an analysis from the global burden of disease study 2010. PloS One. 2015;10(2), e0116820. doi:10.1371/journal.pone.0116820.
46. Kuehner C. Gender differences in unipolar depression: an update of epidemiological findings and possible explanations. Acta Psychiatr Scand. 2003;108(3):163–74.
47. Blazer DG. Depression in late life: review and commentary. J Gerontol Biol Med Sci. 2003;58(3):249–65.
48. Möller-Leimkühler AM. Barriers to help-seeking by men: a review of sociocultural and clinical literature with particular reference to depression. J Affect Disord. 2002;71(1–3):1–9.
49. Richards D. Prevalence and clinical course of depression: a review. Clin Psychol Rev. 2011;31(7):1117–25. doi:10.1016/j.cpr.2011.07.004.
50. Böhnke JR, Croudace TJ. Factors of psychological distress: clinical value, measurement substance, and methodological artefacts. Soc Psychiatr Psychiatr Epidemiol. 2015. doi:10.1007/s00127-015-1022-5.
51. Beck AT, Steer RA, Brown GK. Manual for the Beck Depression Inventory-II. San Antonio, TX: Psychological Corporation; 1996.

52. Wirth RJ, Edwards MC. Item factor analysis: current approaches and future directions. Psychol Meth. 2007;12(1):58–79. doi:10.1037/1082-989x.12.1.58.
53. American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-IV. 4th ed. Washington, DC: American Psychiatric Association; 1994.
54. Lorenzo-Seva U, Ferrando PJ. FACTOR: a computer program to fit the exploratory factor analysis model. Behav Res Meth. 2006;38(1):88–91.
55. Buja A, Eyuboglu N. Remarks on Parallel Analysis. Multivariate Behav Res. 1992;27(4):509–40. doi:10.1207/s15327906mbr2704_2.
56. Timmerman ME, Lorenzo-Seva U. Dimensionality assessment of ordered polytomous items with parallel analysis. Psychol Meth. 2011;16(2):209–20. doi:10.1037/a0023353.
57. Gaskin CJ, Happell B. On exploratory factor analysis: a review of recent evidence, an assessment of current practice, and recommendations for future use. Int J Nurs Stud. 2014;51(3):511–21. doi:10.1016/j.ijnurstu.2013.10.005.
58. Crane PK, Gibbons LE, Ocepek-Welikson K, Cook K, Cella D, Narasimhalu K, et al. A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. Qual Life Res. 2007;16 Suppl 1:69–84. doi:10.1007/s11136-007-9185-5.
59. Scott NW, Fayers PM, Aaronson NK, Bottomley A, de Graeff A, Groenvold M, et al. Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. Health Qualf Life Outcome. 2010;8:81. doi:10.1186/1477-7525-8-81.
60. StataCorp. Stata Statistical Software: Release 12. College Station, TX: StataCorp LP; 2009.
61. Krause JS, Bombardier C, Carter RE. Assessment of depressive symptoms during inpatient rehabilitation for spinal cord injury: Is there an underlying somatic factor when using the PHQ? Rehabil Psychol. 2008;53(4):513–20. doi:10.1037/a0013354.
62. Richardson EJ, Richards JS. Factor structure of the PHQ-9 screen for depression across time since injury among persons with spinal cord injury. Rehabil Psychol. 2008;53(2):243–9. doi:10.1037/0090-5550.53.2.243.
63. Silverstein B. Gender differences in the prevalence of somatic versus pure depression: a replication. Am J Psychiatr. 2002;159(6):1051–2.
64. Kessler RC, Birnbaum H, Bromet E, Hwang I, Sampson N, Shahly V. Age differences in major depression: results from the National Comorbidity Survey Replication (NCS-R). Psychol Med. 2010;40(2):225–37. doi:10.1017/s0033291709990213.
65. Cameron IM, Crawford JR, Lawton K, Reid IC. Differential item functioning of the HADS and PHQ-9: an investigation of age, gender and educational background in a clinical UK primary care sample. J Affect Disord. 2013;147(1–3):262–8. doi:10.1016/j.jad.2012.11.015.
66. Chen H, Cohen P, Chen S. How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. Comm Stat Simulat Comput. 2010;39(4):860–4. doi:10.1080/03610911003650383.
67. Gelin MN, Zumbo BD. Differential item functioning results may change depending on how an item is scored: an illustration with the center for epidemiologic studies depression scale. Educ Psychol Meas. 2003;63(1):65–74. doi:10.1177/0013164402239317.