BMC Psychiatry

Research article

Open Access

Computerized adaptive measurement of depression: A simulation study

William Gardner*[1], Katherine Shear[2], Kelly J Kelleher[1], Kathleen A Pajer[1], Oommen Mammen[2], Daniel Buysse[2] and Ellen Frank[2]

Address: [1]Pediatrics, Children's Research Institute and Ohio State University, Columbus, OH, USA and [2]Psychiatry, Western Psychiatric Institute and University of Pittsburgh, Pittsburgh, PA, USA

Email: William Gardner* - gardnerw@pediatrics.ohio-state.edu; Katherine Shear - shearmk@msx.upmc.edu; Kelly J Kelleher - kellehek@pediatrics.ohio-state.edu; Kathleen A Pajer - pajerk@pediatrics.ohio-state.edu; Oommen Mammen - mammenok@msx.upmc.edu; Daniel Buysse - buyssedj@msx.upmc.edu; Ellen Frank - franke@msx.upmc.edu

* Corresponding author

Published: 06 May 2004

BMC Psychiatry 2004, **4**:13

Received: 23 December 2003
Accepted: 06 May 2004

This article is available from: http://www.biomedcentral.com/1471-244X/4/13

Abstract

Background: Efficient, accurate instruments for measuring depression are increasingly important in clinical practice. We developed a computerized adaptive version of the Beck Depression Inventory (BDI). We examined its efficiency and its usefulness in identifying Major Depressive Episodes (MDE) and in measuring depression severity.

Methods: Subjects were 744 participants in research studies in which each subject completed both the BDI and the SCID. In addition, 285 patients completed the Hamilton Depression Rating Scale.

Results: The adaptive BDI had an AUC as an indicator of a SCID diagnosis of MDE of 88%, equivalent to the full BDI. The adaptive BDI asked fewer questions than the full BDI (5.6 versus 21 items). The adaptive latent depression score correlated *r* = .92 with the BDI total score and the latent depression score correlated more highly with the Hamilton (*r* = .74) than the BDI total score did (*r* = .70).

Conclusions: Adaptive testing for depression may provide greatly increased efficiency without loss of accuracy in identifying MDE or in measuring depression severity.

Background

There is a pressing need for accurate and efficient instruments to screen for depression and to measure its severity, for several reasons. First, the U.S. Preventive Services Task Force [1] recommended adults be screened for depression, based on findings that feedback of depression screening to clinicians increased the recognition of depressive illness. Moreover, a great proportion of depression care is in the hands of clinicians who lack specialized mental health training [2,3]. These clinicians may benefit from methods to detect cases and evaluate the outcomes of care. For example, it has been recognized for some time that depression is an important source of morbidity in primary care [4], and that improvement is needed in the recognition and management of depression in that setting [5,6]. However, clinician time and attention are highly constrained in many health care settings [7].

Second, efficient yet accurate instruments to measure depression would also be of considerable value in monitoring the progress of treatment by mental health specialists and other treating clinicians. Systematic case management activities involving regular outcome measurement are part of an emerging paradigm of high quality care of chronic illnesses [8-11]. However, such programs will fail if patients are unwilling to adhere to measurement protocols. Therefore, follow-up measurement protocols cannot burden patients with repeated exposures to long questionnaires. Finally, researchers frequently assess severity of depression using standard instruments such as the Beck Depression Inventory (BDI) [12]. Efficiency is important here as well, because researchers often assess many constructs, straining study participants' endurance.

Given that many instruments permit assignments of diagnoses, measurement of symptoms, or assessment of functioning, one might ask why routine depression screening and treatment monitoring are not already common. Although these instruments are commonly used in research, they have had less effect on clinical care. One important reason is that the time and other costs of mental health assessments may outweigh their benefits for busy patient-care settings. These costs may also make screening or treatment monitoring suboptimal from a societal perspective. For example, a recent simulation study by Valenstein and her colleagues [13] suggested that screening for depression would not meet a reasonable criterion for cost-utility if the cost of administering a single test was substantially higher than $5, where that cost comprised a fee for the instrument, six minutes of staff time, and one minute of physician time.

Computerized adaptive measurement of depression

Two technical advances could substantially improve the tradeoff between efficiency and accuracy in the measurement of mental health problems, such as depression. First, the Internet is reducing the cost and other barriers to the delivery of computerized testing services to clinical offices. Wireless Internet connectivity is becoming widely available, and powerful, mobile tablet and handheld computers are now available at commodity prices. These technologies should substantially reduce the cost of putting computer-administered tests in the hands of patients and clinicians in front-line clinical settings. Computerized tests, particularly those that can be self-administered by patients, can reduce the staff and clinician time required to administer and score an instrument. There are several computerized mental health instruments including, for example, a computerized version of the Composite International Diagnostic Interview (CIDI) [14,15].

The second technical advance, Computerized Adaptive Testing (CAT) [16], has been widely used by educational and vocational testers, but has seen surprisingly little application in physical or mental health settings [17,18]. CAT is a technology for interactive administration of tests that tailors the test to the examinee (or, in our application, to the patient). These tests are 'adaptive' in the sense that the testing is driven by an algorithm that selects questions in real time and in response to the ongoing responses of the patient. We believe that computerized, adaptive mental health assessment services, delivered on stand-alone computers or over the web, could make significant contributions to both mental health research and clinical care. In this article, we discuss how CAT can be used to screen for, and measure severity of, depression.

The need to achieve both accuracy and efficiency poses a difficult tradeoff for an instrument developer, for two reasons. First, classical test theory [19] teaches that, everything else being equal, the way to make a test more accurate is to increase its length, so that random errors in the responses to individual items cancel each other out.

Second, the need to accurately measure patients with varying levels of severity of disorder lengthens tests. Failing to include items about symptoms reflecting a wide range of severity of disorder will result in an instrument with a floor or ceiling effect [17]. Thus, an accurate and wide-ranging instrument should include several questions at each relevant level of severity of disorder.

Unfortunately, a fixed instrument that has multiple questions for each of several ranges of severity is an inefficient instrument for any individual patient. That individual patient has a disorder the severity of which falls into only one of those ranges, and questions that ask about much more or much less severe symptoms are often irrelevant to that patient. In summary, until recently the goals of having a brief, efficient instrument and an accurate, wide-ranging instrument have seemed mutually incompatible.

Computerized adaptive testing

CAT can improve the terms on which accuracy and efficiency are traded off. It has two components. First, one administers the instrument via computer, using a device such as a touch screen, or through a computer-administered telephone interview [20,21]. Research on computerized tests [22] has shown that the medium has few negative effects on how subjects respond. To the contrary, computerized data collection directly from patients appears to reduce social desirability bias in the reporting of alcohol and drug use, sexual activity, and medication noncompliance [23]. Of particular interest, is the suggestion that people seem to prefer revealing some types of very personal information e.g., gynecological details [24], sexual abuse [25], or suicidal ideation [26] to a computer than a person. Similarly, alcoholics seeking treatment disclosed greater levels of consumption of alcohol to a computer than to a person [27].

CAT, however, goes farther. 'Adaptive' means that the computer follows an algorithm that administers a test (for example, the BDI) to a patient one question at a time. At each step, the patient's prior responses determine (a) whether to ask another question and (b) which question to ask [16,28]. The test stops when the patient's score has been estimated to a prescribed level of precision. Hence, the computer adapts the test to use the fewest items required to assess *that particular patient* accurately. By comparison, an instrument using a fixed list of items may have too few items to accurately measure some patients, while posing unnecessary questions to others.

To that end, at each step, the program uses the current subset of responses to estimate the patient's score on a latent trait, in this case depression, as well as a confidence interval (CI) around that estimate [29]. The latent trait is conventionally denoted as θ, and is conventionally expressed in standardized units. However, θ could be rescaled to the same units as the BDI to aid clinicians familiar with that instrument. The CI around θ is then compared to the 'cut' or criterion score on the latent trait that defines a positive screening result. If the upper bound of the CI were to fall below the cut score, the program would declare the screening result negative, and stop testing. Conversely, if the lower bound of the CI were to fall above the cut score, testing would stop with a positive result. Otherwise, the CI includes the cut score, and testing continues.

Suppose now that we are in mid-test, and the adaptive algorithm has to choose another question to pose to a respondent. Using the data already collected about the respondent, the program calculates an information statistic for each of the test items that have not yet been posed. The information statistic for an item is larger if the response to that item is expected to make a greater reduction in our uncertainty about the patient's true score on the latent depression dimension. The computer then presents the maximally informative item to the respondent. Everything else being equal, a question will be more informative if the severity of the symptoms it concerns is similar to our current estimate of the severity of the patient's depression. For example, if we already have substantial evidence of depression based on the responses thus far, the computer will discount the value of items that primarily ask about minor symptoms, and focus on those that ask about severe symptoms. Please notice that the adaptive algorithm we describe here is different from the branching logic used in many computerized tests to skip questions based on earlier patient responses. Programs that use branching identify questions to be skipped because those questions are irrelevant based on a patient's previous answers. Adaptive tests choose questions to be asked because those questions maximize the precision of the patient's estimated score on a latent dimension of interest.

We reasoned that adaptive technology could substantially improve the efficiency of psychometric measurement in clinical settings, with little or no cost in the accuracy of measurement. We sought to test this by developing an adaptive version of the BDI. We chose the BDI because it is a well-validated instrument for depression and representative of the many screening instruments available for this common condition. It is brief, has been very widely used, and is already in a self-report format.

The goals of this study were (a) to test whether an adaptive version of the BDI would predict a Structured Clinical Interview for DSM-IV Axis I Disorders (SCID) [30] diagnosis as accurately as the full BDI, (b) to estimate how many fewer questions the adaptive BDI would ask, and (c) to determine whether the adaptive BDI would measure the severity of depression as well as the full BDI. The statistical methodology underlying adaptive testing is well established and there is considerable experience in using it in other domains of measurement [16,31]. In a previous study [32], we showed that the screening decisions made by an adaptive version of the Pediatric Symptom Checklist (PSC) [33] agreed nearly perfectly with the screening decisions made by the full PSC (κ = .97). The adaptive PSC achieved that agreement by asking an average of only 10.5 questions per patient, compared to the 35 items required by the full PSC. However, that study did not examine whether adaptive testing affected the PSC's accuracy, which would have required comparing screening decisions based on adaptive data to independent psychometric criteria. To our knowledge, there have been no studies of how an adaptive implementation of a screen for mental health problems affects the agreement between the screen and criterion measures. In this study, we evaluated the performance of an adaptive version of the BDI against an independent SCID diagnosis and Hamilton depression measure.

Methods

Study group and data

This study combined data from nine projects at the University of Pittsburgh. We looked for recent studies in which subjects had both a BDI and a SCID. 1) Two-hundred and nineteen assessments were obtained from mothers seeking treatment for their children in a rural mental clinic from 1998 to 2000. 2) Seventeen depressed women were recruited from a rural mental center in Western Pennsylvania in 1999 for a pilot psychotherapy protocol. 3) Twenty-three subjects participated in a pilot study of cognitive behavioral treatment for traumatic grief in 1999 and 2000. 4) Forty-three women came from a descriptive study of anger in pregnant or post-partum women [34] who presented for treatment of mood and anxiety disorders in a psychiatric clinic in a university medical center in 1996 and 1997. 5) Eighty-seven subjects came from a study of maintenance therapy in bipolar disorder. 6) Nine subjects came from a study of borderline personality disorder. 7) Fourteen subjects came from a pilot study of brief interpersonal psychotherapy. 8) One hundred eighty-three subjects came from a study of maintenance psychotherapy in women with recurrent major depression. 9) Finally, 149 subjects came from a study of normal sleeping patterns in adults. These latter subjects were selected based on having no lifetime history of mental disorders as measured by the SADS or the SCID, as well as no first-degree family history of mental disorders.

For 285 of these patients, we also had Hamilton Depression Rating Scale [35] scores obtained within one week of the BDIs to serve as an independent measure of the severity of depression. For this subset, we were able to compare whether the Adaptive BDI correlated with the Hamilton as well as the total score of the full BDI.

Pooling these data sets resulted in 744 subjects. Of these, 84% were female, 91% were European-Americans, and the average age was 37 (*SD* = 8.6 years). All subjects in these studies had completed a BDI and had received a diagnostic evaluation with the SCID. Patients completed the BDI before treatment, during a symptomatic period at or near the time of the diagnostic interview. Three hundred thirty-nine participants had either a SCID diagnosis of major depressive disorder, or bipolar disorder in which it could be established through independent and concurrent assessments that the patient had completed the BDI in a depressed phase. These unipolar and bipolar depressives were classified as having an MDE. Of the remaining 405 participants, 256 had diagnoses other than MDE, and 149 had no diagnosed disorders.

Beck Depression Inventory (BDI)

The BDI is a widely used 21-item depression survey (there is an additional skip-out item that was ignored in this analysis). Each item on the BDI includes four response statements describing increasing severity of depression. A few (<0.2%) scores on specific BDI questions were missing. Randomly imputed scores replaced these values.

Item Response Theory (IRT) modeling

IRT [36-38] has replaced classical test theory [19] as the leading psychometric theory for surveys and tests in education, the social sciences, and increasingly, for patient-reported data in health care [17,18,32]. In a test created using classical test theory, there are points assigned to each response to each question (for example, 1 point for a 'yes' response to a yes/no question about a depression symptom and 0 points for a 'no' response). You would score the test by summing the points to compute a total score. You would interpret the result by locating that total score to the distribution of total scores in a normative sample, perhaps judging the result problematic if the total score fell in the upper 10% of a national sample of respondents. IRT is based on a mathematical model that, for each item on a test, regresses the person's response to the item on a latent score that represents the attribute of the person that the instrument measures. The person's score on the test is estimate of the value of the latent variable that maximizes the likelihood of the person's pattern of responses. In the proposed research, the latent dimension of interest might be viewed as the severity of the patient's substance use. To model the BDI, we used the graded-response model [36], a variant of IRT for polytomous data.

Adaptive testing simulation

The goal of this study was to determine how well the adaptive BDI predicted SCID diagnoses of MDE, how well it measured depression severity, and how efficient it was compared to the regular BDI. To simulate the adaptive use of the BDI, we wrote a program that interacted with the Adaptive BDI. This program simulated a patient taking the test by using participants' paper and pencil BDI data as if they had been collected adaptively. For each participant, the simulation began by asking the question that was most informative based on the assumption that the participant's latent depression score was the population mean; this is the BDI's question 7, which concerns the subject's disappointment with self. However, we knew how each participant had responded to question 7 on the paper and pencil BDI, and we assumed that he or she would have made the same response if the question had been asked through an adaptive process. Taking the participant's actual response to question 7 as the response to the first question in the simulated adaptive testing session, the computer used the adaptive algorithm to choose the next question. Similarly, at each subsequent step we used the participants' actual responses to drive the algorithm forward.

Next, we used the simulated patient program to measure how well the Adaptive BDI would predict MDE and the BDI total score, using the strategy of internal cross-validation [39]. In this strategy, we first partitioned the 744 cases into 100 groups of seven or eight participants. We then held out the cases in the first of the hundred groups and estimated the IRT model underlying the adaptive BDI using the remaining 99% of the data. The parameter estimates from the IRT model were then substituted into the program implementing the Adaptive BDI. The patient simulation program then used the 1% of participants whose data had not been used in the IRT estimation in a simulation of the adaptive use of the BDI. We then repeated this procedure for each of the other 1% subgroups of participants, until all 744 participants had served as 'fresh' cases in the simulations of the adaptive use of the BDI.

To compute an ROC curve for the adaptive BDI, we evaluated how the adaptive BDI would behave at each of 30 evenly spaced cut points on θ (the latent depression score), ranging from -4.0 to 4.0 (θ has mean 0 and standard deviation 1). Through simulation, for each cut point we determined how many questions the algorithm would ask for each participant, and what screening decision it would make when it stopped. Thus, it was possible to compute the sensitivity and specificity of the adaptive BDI for each θ cut point, and therefore to calculate its ROC curve and the AUC.

To measure the efficiency of the adaptive BDI, we calculated the average number of questions asked by the adaptive algorithm at a cutpoint that offered high levels of both sensitivity and specificity. To assess how well it functioned as a measure of depression severity, we calculated the correlation between the adaptive BDI and the Hamilton scale.

Simulating variance in the prevalence of MDE

An important problem in our study was that we used research samples, in which the prevalence of MDE was higher than would be found in many clinical settings. To address this, we compared the adaptive BDI and the regular BDI in several bootstrapping analyses [40] in which we oversampled cases that did not have a diagnosis of MDE.

Results

IRT analysis of the BDI

The transformation of an existing instrument into an adaptive test begins with a psychometric analysis of the instrument, based on Item Response Theory (IRT) [18,36,41,42]. To this end, we first performed a factor analysis of the BDI data to assess the dimensionality of the instrument. Unidimensionality of the factor structure of the test items – which means that the associations among patients' responses to the BDI items can be accounted for by a single factor – is an important assumption underlying unidimensional IRT and CAT [36]. We used a factor analysis model appropriate for ordinal categorical data [43], and estimated it using the program Mplus [44].

In our factor analysis, the first factor accounted for 58% of the variance in the BDI (eigenvalue = 12.2), while the next factor accounted for 6% (eigenvalue = 1.2). Fitting a one-factor model to the data produced a root mean-square residual statistic of .048. This statistic ranges between 0 and 1, with small values reflecting a better fit; .05 is often used as a criterion for adequacy of fit. We concluded that a unidimensional model fit the data adequately, as did Clark and his colleagues [45]. Other authors have fit more than one correlated factor to different sets of BDI data [46-48]. Differing results in factor analyses often reflect differences in sample selection. Our study involved a mixture of patients and healthy participants and it is likely that there was greater variance in the severity of depression among these patients than in studies including primarily psychiatric cases or primarily healthy participants such as college students. If so, we would expect to find a large first factor measuring severity of depression that accounted for a high proportion of the variance in BDI responses. Having established that a unidimensional solution fit the data, the IRT modeling was performed using the program PARSCALE [49,50].

ROC analysis of the Adaptive BDI

The baseline for evaluating the accuracy of the adaptive BDI was the accuracy of the 21-item BDI, so we began by computing the Area Under the Curve (AUC) of the ROC curve [51] for the BDI total score ($\bar{X}$ = 16.3, SD = 12.8), when the latter was used as an indicator of a SCID diagnosis of MDE. The ROC curve for the 21-item BDI total score had an AUC = 89.4% (95% confidence interval = [87.1%, 91.7%]). The ROC curve for the adaptive BDI (Figure 1) was almost identical, with AUC = 88.4%. Note that this statistic and all the following results are cross-validated estimates.

We then examined the ROC curve for the adaptive BDI and chose the point that offered the best combination of high sensitivity and high specificity (sensitivity = 87.6%, specificity = 79.3%, positive predictive value = 78.0%, negative predictive value = 88.4%; cases were judged to be positive if $\theta \geq .135$; this point is labeled 'Best Case' in Figure 1). Table 1 presents the agreement between the adaptive BDI and the SCID for this case. Figure 2 presents the distributions of estimated θ scores, depending on whether the patient had no diagnoses, a depression diagnosis other than MDE, or MDE. The Kappa [52] for BDI-SCID concordance was .66, which is considered a good level of agreement by conventional standards. The average number of questions asked was 5.6 (SD = 6.6), and for 69% of the subjects the algorithm asked fewer than five questions (Figure 3). In addition, the algorithm asked more questions about cases in which it decided that the participant had an MDE ($\bar{X}$ = 6.3, SD = 7.2) than in cases in which it decided that an MDE was not present ($\bar{X}$ = 4.9, SD = 5.8; Levene's $t(725) = 2.83$, $p < .005$).

Finally, we asked whether the adaptive BDI would be as useful as the full BDI as a measure of the severity of

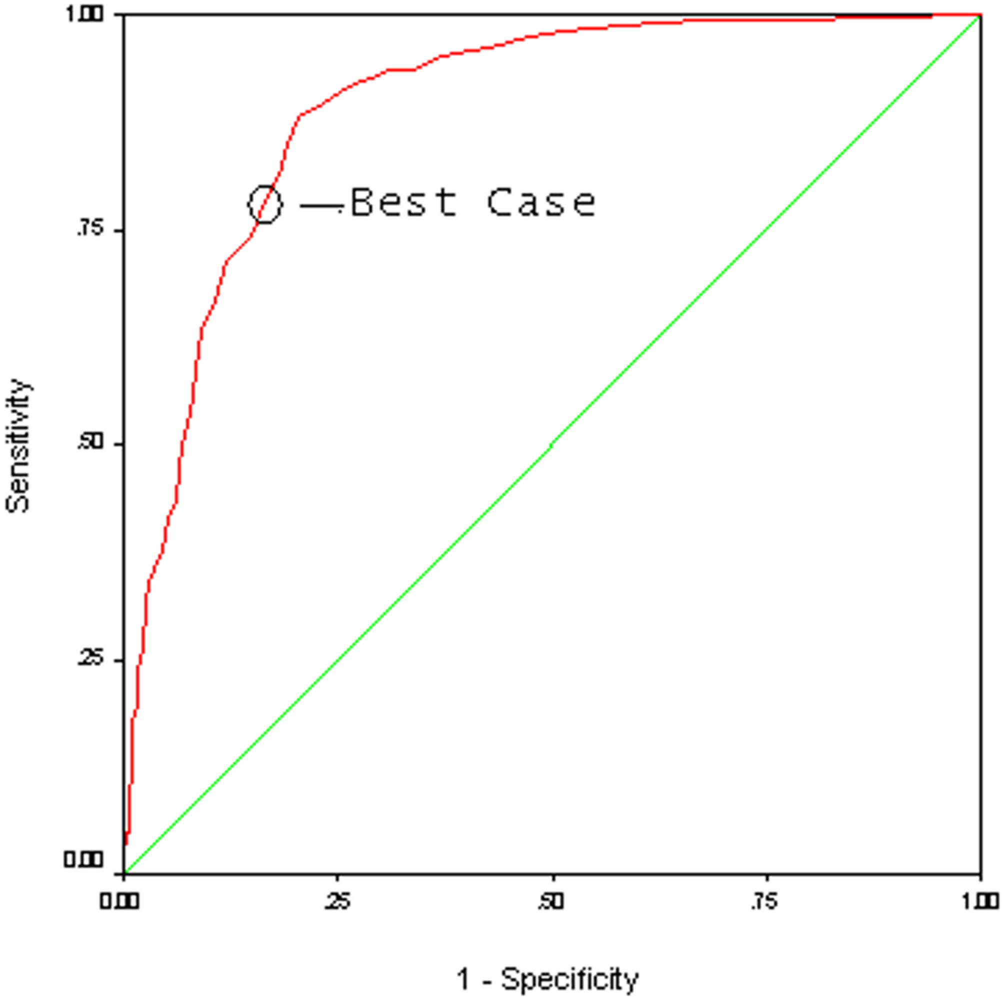

Figure 1
ROC curve for Adaptive BDI as an indicator of Major Depressive Disorder

Table 1: Cross-validated agreement between adaptive BDI 'best case' and SCID Major Depressive Episode: Unweighted results

	SCID		
Adaptive BDI	Negative	Positive	Total
Negative	321	42	363
Positive	84	297	381
Total	405	339	744

depression. The estimated latent depression scores ($\hat{\theta}$) in the 'best case' simulation were highly correlated (r = .92, N = 744) with the BDI total score. For the 285 clinical cases for whom we had both a BDI and a Hamilton score,

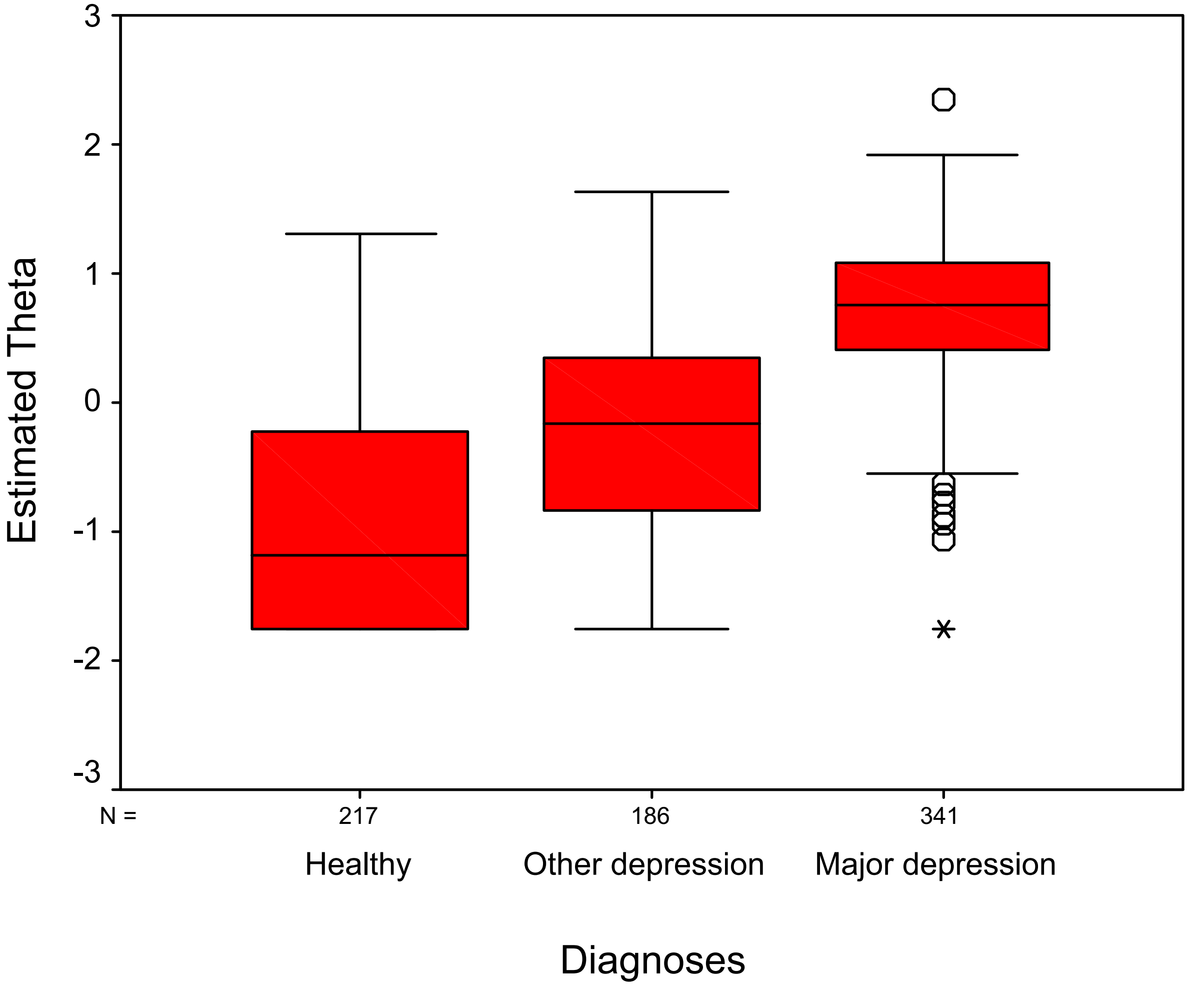

Figure 2
Box plot of distribution of Θ by depression diagnosis

the BDI total score had a correlation of $r = .70$ with the Hamilton, while the correlation between $\hat{\theta}$ and the Hamilton was $r = .74$. This difference is statistically significant [p < .006, [53]].

The effect of prevalence of MDE

Our study group included more positive cases than a sample that might be found in medical settings other than specialty mental health settings. To examine whether our results would still hold if the prevalence of MDE were lower, we conducted additional simulations in which we created new study groups of cases by randomly sampling cases with replacement from our data (i.e., bootstrapping). We generated 1000 bootstrap samples in which the sampling weights on cases were set such that the average prevalence of MDE in the bootstrap samples was 10%. We then repeated our AUC analyses in each bootstrapped sample. The results suggested that both the regular BDI and adaptive BDI performed as well or better when the prevalence of MDE was lower. That is, the average AUC for the agreements between the adaptive BDI and the SCID was 92.4%, and the average AUC for the agreements between the regular BDI and the SCID was 92.3%.

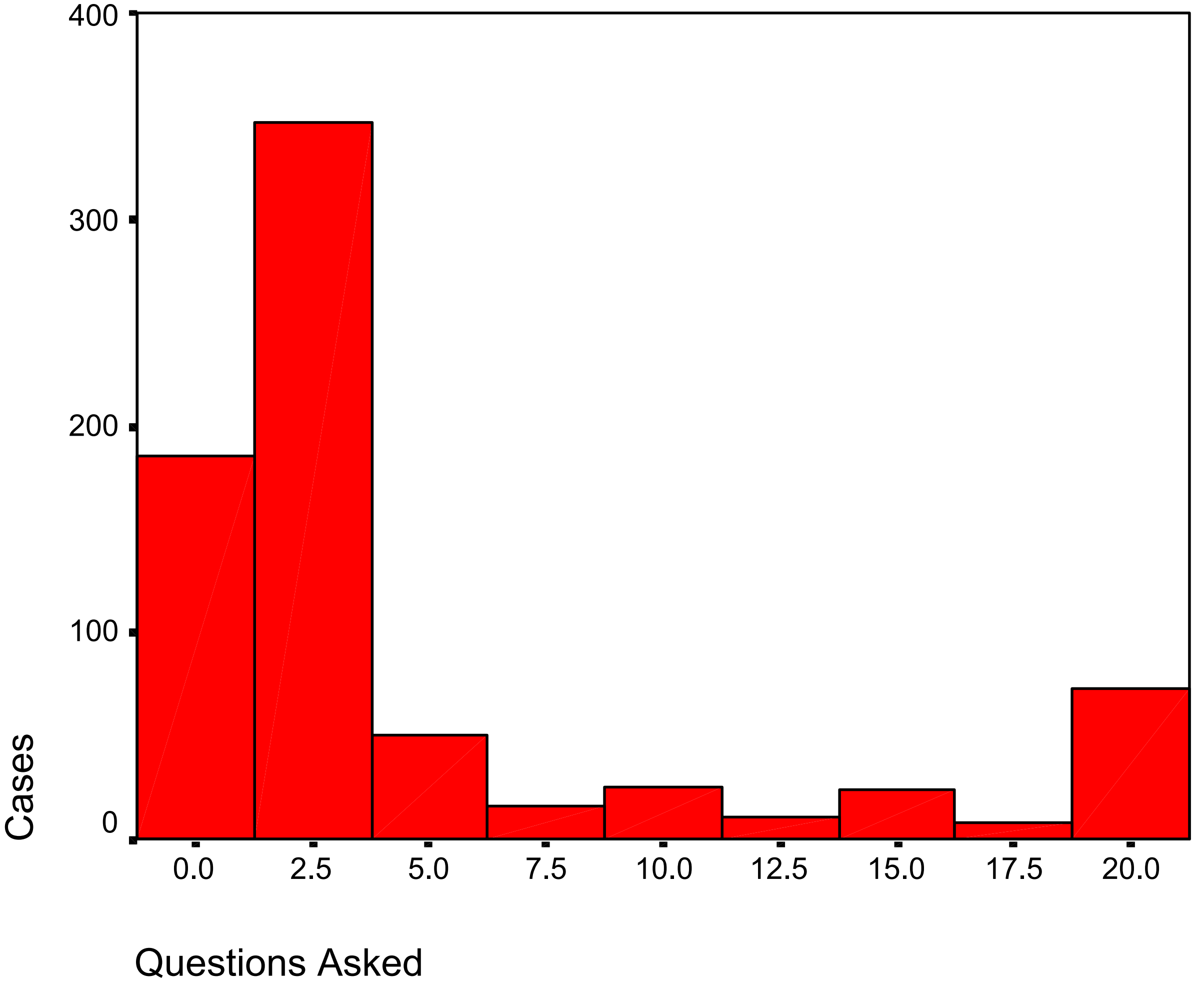

Figure 3
Histogram of questions asked in 'best case' simulation **Questions**

Discussion

The AUC for the adaptive BDI was a respectable 88%, indicating that the adaptive test could correctly classify large proportions of both positive and negative cases. The 'best case' adaptive BDI was able to classify a subject using an average of only 5.6 questions. The latent depression score generated in that simulation was highly correlated with the BDI total score. In addition, for the subset of the data for which Hamilton scores were available, the latent score was more highly correlated with the Hamilton than the BDI total score was. The latter results indicate that the adaptive BDI would be as useful as the full BDI as a measure of the severity of depression. Thus, in our simulation the adaptive BDI was as accurate as the full scale BDI while dramatically improving efficiency. We note that the CAT algorithm can be 'tuned' to the assessment purpose at hand, for which one might choose another point that emphasized either sensitivity or specificity.

The results also suggested, however, that five or six questions would be the required number of items for only a few participants. Indeed, the adaptive BDI asked fewer than five questions for the majority of patients. Even when the adaptive BDI asked few questions, it usually made the same screening decision as the full BDI: the rates of disagreements between the adaptive BDI and the SCID were slightly higher when more questions were asked. In addition, the algorithm asked more questions about positive cases. This is an attractive outcome, because the patients' answers provide useful symptom data for the clinician. Thus, an adaptive test budgets the patient and clinician time spent on measuring depression, allocating it primarily to persons for whom there are reasons for concern.

A reduction of 15 questions may not seem important, given that the full BDI takes only a few minutes to complete. In our experience, however, clinicians are very concerned about both office visit time and maintaining a smooth flow of patients through the waiting room. In addition, health care providers are confronted with recommendations that they screen for many illnesses and health-related problems. Saving questions on a depression screen might free time to screen for other problems such as domestic violence or substance abuse.

Based on this simulation, it appears that adaptive testing has significant promise for settings where both high efficiency and high accuracy are essential. These settings include primary care, where clinician time is a rate-limiting factor, and ongoing monitoring after successful treatment in specialty mental health care, where respondent burden is a constraint. The next step should be to field test adaptive instruments and validate them prospectively, to make certain that they are acceptable to patients and clinicians, and to measure the costs of implementing them.

Limitations

The principal limitation of our study is that it is a simulation. In particular, we assumed that participants would respond to questions presented adaptively similarly to the way they responded on the paper BDI, and that the order in which questions are asked does not have an important effect on accuracy. We have some confidence in these assumptions, based on prior research showing that adaptive versions of tests in other fields have accuracy that is similar to paper and pencil versions [16].

A second limitation is that we compiled our study group from several research studies. Although the study group included a wide range of both healthy and acutely ill individuals, it would be preferable to have a random sample from a defined health care setting. We attempted to statistically control the prevalence of MDE in our bootstrap analyses and found that the performance of the regular and adaptive BDI improved slightly when the prevalence of MDE was decreased. While this is reassuring, we speculate that because our study group may have lacked some of the mildly depressed and difficult to classify cases that are found in many real world settings. The sample that we have collected may also explain why we found evidence for only one factor in the BDI data, where other researchers have found evidence for two or more. Although we view our study as providing strong support for the *concept* of adaptive measurement of depression, because of these limitations it needs to be replicated in an actual sample.

A final limitation of our study is the BDI itself. Our results argue that a computerized adaptive test has the same sensitivity and specificity as its full-length version, and the same validity as a measure of symptom severity. Conversion to CAT cannot, however, produce a test with *higher* validity than the base instrument. Although the BDI is widely used, one may still judge that, in light of a false positive rate of 21%, neither the regular nor the adaptive BDI is sufficiently accurate to serve as a screen for MDE. Our results nevertheless suggest that whatever depression instrument one chooses, it would be more efficient to administer it adaptively.

Conclusions

We believe that adaptive testing could substantially improve the tradeoff between accuracy and efficiency in the assessment of psychopathology. The simulation conducted here showed that the BDI, a widely used depression instrument, could be converted to a far more efficient adaptive test without loss of accuracy. We need more studies to assess the performance of adaptive tests in both mental health specialty and other clinical settings.

Competing interests

None declared.

Authors' contributions

William Gardner: Principal author, data analyst, computer programming, and conceptualization of the study.

Katherine Shear: Conceptualization of study, contribution of data, expertise on psychiatric measurement, critical revision of text.

Kelly J. Kelleher: Conceptualization of study, critical revision of text.

Kathleen A. Pajer: Conceptualization of study, expertise on psychiatric measurement, critical revision of text.

Oommen Mammen: Contribution of data, expertise on psychiatric measurement, critical revision of text.

Daniel Buysse: Contribution of data, expertise on psychiatric measurement, critical revision of text.

Ellen Frank: Contribution of data, expertise on psychiatric measurement, critical revision of text.

Acknowledgements

This study was supported by grants from the NIMH (MH30915, MH29618, MH49115, MH53817, MH56848) and the Staunton Farm Foundation. We thank Drs. David Kupfer and Victoria Grochocinski for comments. Thanks to Mary McShea, and Deb Stapf for assistance in assembling the data.

References

1. U.S. Preventive Services Task Force: **Guide to Clinical Preventive Services. Screening: Depression.** 3rd Edition: Periodic Updates2002, **2002:** [http://www.ahcpr.gov/clinic/uspstf/uspsdepr.htm]. Agency for Health Care Research and Quality
2. Schurman RA, Kramer PD, Mitchell JB: **The hidden mental health network.** *Archives of General Psychiatry* 1985, **42:**89-94.
3. Regier DA, Narrow WE, Rae DS, Manderscheid RW, Locke BZ, Goodwin FK: **The de facto US mental and addictive disorders service system: Epidemiologic Catchment Area prospective 1-year prevalence rates of disorders and services.** *Archives of General Psychiatry* 1993, **50:**85-94.
4. Katon W, Schulberg H: **Epidemiology of depression in primary care.** *General Hospital Psychiatry* 1992, **14:**237-247.
5. Simon GE, VonKorff M: **Recognition, management, and outcomes of depression in primary care.** *Archives of Family Medicine* 1995, **4:**99-105.
6. Schulberg HC, Katon WJ, Simon GE, Rush J: **Best clinical practice: Guidelines for managing major depression in primary medical care.** *Journal of Clinical Psychiatry* 1999, **60:**19-26.
7. Mechanic D, McAlpine DD, Rosenthal M: **Are patients' office visits with physicians getting shorter?** *New England Journal of Medicine* 2001, **344:**223-225.
8. Von Korff M, Gruman J, Schaefer JK, Curry SJ, Wagner EH: **Collaborative management of chronic illness.** *Annals of Internal Medicine* 1997, **127:**1097-1102.
9. Wagner EH, Austin BT, Von Korff M: **Improving outcomes in chronic illness.** *Managed Care Quarterly* 1996:12-25.
10. Katon W, Rutter C, Ludman EJ, Von Korff M, Lin E, Simon G, Bush T, Walker E, Unutzer J: **A randomized trial of relapse prevention of depression in primary care.** *Archives of General Psychiatry* 2001, **58:**241-247.
11. Rost K, Nutting P, Smith JL, Elliot CE, Dickinson M: **Managing depression as a chronic disease: A randomised trial of ongoing treatment in primary care.** *British Medical Journal* 2002, **325:**934-939.
12. Beck AT, Steer RA: **Manual for the Beck Depression Inventory.** San Antonio, TX, Psychological Corporation; 1993.
13. Valenstein M, Vijan S, Zeber JE, Boehm K, Buttar A: **The cost-utility of screening for depression in primary care.** *Annals of Internal Medicine* 2001, **134:**345-360.
14. World Health Organization: **CIDI-Auto Version 2.1: Administrator's Guide and Reference.** Sydney, Training and Reference Centre for WHO CIDI; 1997.
15. Peters L, Clark D, Carroll F: **Are computerized interviews equivalent to human interviewers? CIDI-Auto versus CIDI in anxiety and depressive disorders.** *Psychological Medicine* 1998, **28:**893-901.
16. Wainer H: **Computerized adaptive testing: A primer.** 2nd edition. Hillsdale, NJ, Erlbaum Associates; 2000.
17. McHorney CA: **Generic health measurement: Past accomplishments and a measurement paradigm for the 21st century.** *Annals of Internal Medicine* 1997, **127:**743-750.
18. Hays RD, Morales LS, Reise SP: **Item Response Theory and health outcomes measurement in the 21st century.** *Medical Care* 2000, **38:**II-28-II-42.
19. Crocker L, Algina J: **Introduction to classical and modern test theory.** New York, Holt, Rinehart, & Winston; 1986.
20. Leon AC, Kelsey JE, Pleil A, Burgos TL, Potera L, Lowell KN: **An evaluation of a computer assisted telephone interview for screening for mental disorders among primary care patients.** *Journal of Nervous and Mental Diseases* 1999, **187:**308-311.
21. Kobak K, Taylor L, Dottl S, Griest J, Jefferson J, Burroughs D, Mantle J, Katzelnick D, R Norton., Henk J, Serlin R: **A computer-administered telephone interview to identify mental disorders.** *Journal of the American Medical Association* 1997, **278:**905-910.
22. Mead AD, Drasgow F: **Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis.** *Psychological Bulletin* 1993, **114:**449-458.
23. Millstein SG, Irwin CE: **Acceptability of computer-acquired sexual histories in adolescent girls.** *Journal of Pediatrics* 1983, **103:**815-819.
24. Slack WV, Van Cura LJ: **Patient reactions to computer-based medical interviewing.** *Computers in Biomedical Research* 1968, **1:**527-531.
25. Bagley C, Genuis M: **Psychology of computer use: XX. Sexual abuse recalled: Evaluation of a computerized questionnaire in a population of young adult males.** *Perceptual and Motor Skills* 1991, **72:**287-288.
26. Greist JH, Gustafson DH, Strauss FF, Rowse GL, Laughren TP, Chiles JA: **A computer interview for suicide-risk prediction.** *American Journal of Psychiatry* 1973, **130:**1327-1332.
27. Lucas RW, Mullin PJ, Luna CB, McInroy DC: **Psychiatrists and a computer as interrogators of patients with alcohol-related illnesses: A comparison.** *Br J Psychiatry* 1977, **131:**160-167.
28. Revicki DA, Cella DF: **Health status assessment for the twenty-first century: Item response theory, item banking and computerized adaptive testing.** *Quality of Life Research* 1997, **6:**595-600.
29. Bock RD, Mislevy RJ: **Adaptive EAP estimation of ability in a microcomputer environment.** *Applied Psychological Measurement* 1982, **6:**431-444.
30. First MB, Spitzer RL, Gibbon M, Williams JBW: **Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I).** Washington, DC, American Psychiatric Press; 1997.
31. Weiss DJ: **Adaptive testing by computer.** *Journal of Consulting and Clinical Psychology* 1985, **53:**774-789.
32. Gardner W, Kelleher KJ, Pajer KA: **Multidimensional adaptive testing for mental health problems in primary care.** *Medical Care* 2002, **40:**812-823.
33. Jellinek MS, Murphy JM, Robinson J, Feins A, Lamb S, Fenton T: **Pediatric symptom checklist: Screening school-age children for psychosocial dysfunction.** *Journal of Pediatrics* 1988, **112:**201-209.
34. Mammen O, Shear MK, Pilkonis PA, Kolko DJ, Thase MA, Greeno C: **Anger attacks: Correlates and significance of an underrecognized symptom.** *Journal of Clinical Psychiatry* 1999, **60:**633-642.
35. Hamilton M: **Development of a rating scale for primary depressive illness.** *British Journal of Social and Clinical Psychology* 1967, **6:**278-296.
36. Embretson SE, Reise SP: **Item response theory for psychologists.** Mahwah, NJ, Lawrence Erlbaum; 2000.
37. Hambleton RK, Swaminathan H, Rogers HJ: **Fundamentals of item response theory.** Newbury Park, CA, Sage; 1991.
38. van der Linden W, Hambleton RK: **Handbook of modern item response theory.** New York, Springer Verlag; 1997.
39. Efron B: **Estimating the error rate of a prediction rule: Improvements on cross-validation.** *Journal of the American Statistical Association* 1983, **78:**316-331.
40. Efron B: **The jackknife, the bootstrap, and other resampling plans.** Philadelphia, SIAM; 1982.
41. Dodd BG, De Ayala RJ, Koch WR: **Computerized adaptive testing with polytomous items.** *Applied Psychological Measurement* 1995, **19:**5-22.
42. Hambleton RK, Swaminathan H: **Item response theory: Principles and applications.** Boston, Kluwer-Nijhof; 1985.
43. Muthén BO: **A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators.** *Psychometrika* 1984, **49:**115-132.
44. Muthén LK, Muthén BO: **Mplus user's guide.** Los Angeles, CA, Muthen & Muthen; 1998.
45. Clark DC, vonAmmon Cavansugh S, Gibbons RD: **The core symptoms of depression in medical and psychiatric patients.** *J Nerv & Ment Dis* 1983, **171:**705-713.
46. Gibbons RD, Clark DC, Cavanaugh S, Davis JM: **Application of modern psychometric theory in psychiatric research.** *Journal of Psychiatric Research* 1985, **19:**43-55.
47. Tanaka Jeffrey S, Huba George J: **Confirmatory hierarchical factor analyses of psychological distress measures.** *Journal of Personality & Social Psychology* 1984, **46:**621-635.
48. Tanaka Jeffrey S, Huba George J: **Structures of psychological distress: Testing confirmatory hierarchical models.** *Journal of Consulting & Clinical Psychology* 1984, **52:**719-721.
49. Muraki E: **Fitting a polytomous item response model to Likert-type data.** *Applied Psychological Measurement* 1990, **14:**59-71.

50. Muraki E, Bock RD: **PARSCALE: IRT item analysis and test-scoring for rating-scale data.** Chicago, Scientific Software International; 1997.
51. Kraemer HC: **Evaluating medical tests: Objective and quantitative guidelines.** Newbury Park, CA, Sage Publications; 1992.
52. Cohen J: **A coefficient of agreement for nominal scales.** *Educational and Psychological Measurement* 1960, **20:**37-46.
53. Meng X-L, Rosenthal R, Rubin DB: **Comparing correlated correlation coefficients.** *Psychological Bulletin* 1992, **111:**172-175.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-244X/4/13/prepub