Estimating the accuracy of psychological scales using longitudinal data

CAROLYN M. RUTTER*, DIANA L. MIGLIORETTI
Group Health Cooperative, Center for Health Studies, 1730 Minor Ave Ste. 1600, Seattle WA, 98101, USA
rutter.c@ghc.org

SUMMARY
This article presents a method for estimating the accuracy of psychological screening scales using receiver operating characteristic curves and associated statistics. Screening scales are typically semi-continuous within a known range with distributions that are nearly symmetric when the target condition is present and highly skewed when the condition is absent. We model screening scale outcomes using truncated normal distributions that accommodate these different distributional shapes and use subject-specific random effects to adjust for multiple assessments within individuals. Using the proposed model, we estimate the accuracy of the Symptom Checklist as a measure of major depression from a repeatedly screened sample of patients.

Keywords: Cutpoints, Receiver operating characteristic curve, Truncated Normal.

1. INTRODUCTION
Screening scales play an important role in both clinical mental health care and mental health research. In clinical settings, screening scales are part of initial evaluation and treatment follow-up. Screening scales are also used for epidemiologic, effectiveness and efficacy studies. Because psychological screening scales approximate clinical diagnoses, the usefulness of these scales depends on both their ease of administration and their accuracy relative to more complicated ‘gold standard’ assessments. In this paper, we focus on estimating the accuracy of a depression screening test. We also estimate cutpoints on the screening scale that yield dichotomized results with particular characteristics, such as a target sensitivity or positive predictive value.

Estimating the accuracy of screening scales can be difficult because the distributions of test results are often very different in affected and unaffected populations. Screening scales generally include an inventory of symptoms, with scale scores increasing with both the number and severity of symptoms. Because these scales are bounded, they are subject to both floor and ceiling effects. Floor effects occur when unaffected individuals are completely asymptomatic. As a result, scale scores tend to be right-skewed in unaffected populations. Ceiling effects can similarly result in left-skewness of the scale score distribution among severely affected populations, though scale scores often have an approximately symmetric distribution among affected individuals.

We present a parametric method for estimating the accuracy of a screening scale that allows differences in distributional shapes for affected and unaffected populations. Our approach estimates the accuracy

*To whom correspondence should be addressed

of test outcomes using receiver operating characteristic (ROC) curves estimated from screening scale distributions. This work was motivated by an examination of the accuracy of and cutpoints for the Hopkins Symptom Checklist (SCL) using longitudinal data from a randomized clinical trial (RCT) of antidepressant treatment. We use a Bayesian modeling approach to account for correlation due to the multiple assessments within individuals.

In the next section we describe the RCT data and the goals of our analysis. We review key aspects of ROC analysis in Section 3 and introduce our modeling approach in Section 4. Finally, we use the RCT data to estimate population-average estimates of ROC curves and optimal cutpoints on the SCL scale with 95% credible intervals. We conclude with summary remarks.

2. EXAMPLE: RANDOMIZED TRIAL OF INITIAL ANTIDEPRESSANT PRESCRIPTION

Our example comes from a randomized trial conducted to examine differences in the long-term clinical outcomes, quality of life, and health care costs for patients initially prescribed one of three antidepressants (Simon et al., 1996). Patients who were about to begin antidepressant treatment were referred to the study by their primary care provider. 536 patients were randomized to receive an initial prescription for fluoxetine, imipramine, or desipramine. Neither patients nor their physicians were blinded to the treatment they received, and patients were allowed to switch medication. Patients were interviewed by phone at baseline and 1, 3, 6, 9, 12, 18, and 24 months following the initiation of treatment. At each interview, the patient’s depression status was assessed using both the SCL and the Depression module of the Structured Clinical Interview (SCID). The SCID provides an assessment of current major depression in accordance with the Diagnostic and Statistical Manual (DSM) and is the traditional ‘gold standard’ for depression assessment. The SCID assesses the presence and duration of nine depression symptoms. By DSM-IV criteria, patients with major depression experience at least one cardinal symptom (loss of interest or depressed mood) and a total of five out nine possible symptoms, with all of these symptoms continually present for at least 2 weeks.

While the SCID is a well accepted research tool for diagnosing depression, it is not useful as a screening instrument. The SCID is administered by a trained (lay) interviewer and is not readily adapted to self-administration. Therefore, the SCID cannot be used as part of a mailed interview, or as part of a self-assessment that can be administered before a primary care visit. In addition, the SCID captures only the presence of symptoms over a 2 week period and does not capture symptom intensity. In contrast, the SCL provides a quick and easy to self-administer instrument for measuring depression symptom severity, but does not provide a clinical diagnosis of depression. The SCL asks respondents to rate how much they were ‘distressed’ by symptoms of depression during the last month, with each symptom rated on a scale from zero (not at all) to four (extremely). The modified version of the SCL administered as part of this randomized trial includes 20 items that are averaged to give a depression severity scale that ranges from 0 to 4. This trial found no differences in depression outcomes for the three treatment groups.

Table 1 shows the distribution of the number of follow-up assessments across patients. Most patients (95% = 509/536) had at least one follow-up assessment, and 71% (379/536) completed the 24 month interview. These data provide an opportunity to examine the accuracy of the SCL as a screen for major depression following pharmacotherapy. The SCL is a convenient research instrument and may be a useful clinical tool for measuring of response to treatment. However, the ability of SCL to detect SCID-identified depression is not known. In addition, the clinical utility of the SCL has been limited, because appropriate cutpoints for screening have not been determined. In the following sections we present methods and analyses that use these longitudinal data to estimate (1) the distribution of SCL scores for depressed and not-depressed individuals, (2) the accuracy of the SCL as a screen for major depression following treatment, and (3) optimal cutpoints on the SCL scale for determining screen positive and screen negative status.
3. RECIIVER OPERATING CHARACTERISTIC ANALYSIS

ROC analysis is used to estimate the accuracy of ordinal and continuous tests relative to a dichotomous true disease state \((D = 1\) when diseased and \(D = 0\) when not-diseased). ROC analysis focuses on positive test outcomes: true positive rates (TP) and false positive rates (FP). The TP rate is the proportion with disease who test positive. The FP rate is the proportion without disease who test positive. For a dichotomous test, there is one \((TP, FP)\) pair and this pair can be used to summarize test accuracy. When a test outcome is ordinal or continuous, multiple \((TP, FP)\) pairs are possible and ROC analysis combines these pairs to estimate overall test accuracy. We focus on estimates of overall accuracy for monotonic tests and assume that the probability of disease is a non-decreasing function of test outcome, \(Y\). For an arbitrary cutpoint, \(c\), individuals are classified as positive when \(Y \geq c\). For example, we can define a dichotomous test that is positive for depression when individuals score 0.75 or higher on the SCL. Given \(c\), the TP rate of a test is \(P(Y \geq c \mid D = 1) = S_{D}(c)\) and the FP rate is \(P(Y \geq c \mid D = 0) = S_{\bar{D}}(c)\).

The ROC curve describes the multiple \((TP, FP)\) pairs corresponding to all possible cutpoints for a single ordinal or continuous diagnostic test. The empirical ROC curve is based on observed \((TP, FP)\) pairs. Figure 1 shows empirical ROC curves for the SCL as a screen for SCID-identified depression across all time points. A perfect test has FP = 0 and TP = 1, and a chance test has FP equal to TP. As test accuracy increases, the ROC curve moves away from the 45° line (TP = FP) and towards (0, 1).

In this paper, we develop a method for estimating ROC curves and associated statistics when test outcomes are continuous and have skewed distributions and data are clustered within individuals. There are at least three general ways of approaching this problem: bootstrap of empirical ROC curves, use of generalized estimating equations for regression models, and random effects models that are typically fit using a Bayesian framework.
The simplest way to estimate a single ROC curve from clustered data is by bootstrapping the empirical ROC curve. The empirical ROC curve is given by $\text{ROC}(FP) = \hat{S}_D(\hat{S}^{-1}_D(FP))$ where $\hat{S}_D(\cdot)$ and $\hat{S}_D(\cdot)$ are empirical survivor functions for test results in the diseased and not diseased groups. Using the bootstrap approach, clusters (e.g. individuals) are resampled with replacement and for each bootstrap sample, $(TP, FP)$ pairs are calculated for each cutpoint on the test scale. The overall ROC curve is estimated by averaging the $(TP, FP)$ pairs for each cutpoint across the bootstrap samples.

Generalized estimating equations have been proposed for both ordinal (Toledano and Gatsonis, 1996) and continuous (Pepe, 1997, 1998, 2000) test outcomes. The ordinal regression model for ROC estimation is given by $g(P(Y \geq c | D)) = (\theta_c + \alpha D)e^{-\beta D}$ (Hanley, 1989). Under this model, $\theta_c$ estimates overall rate of test positivity, $\alpha$ estimates test accuracy through differences in TP and FP rates, and $\beta$ is used to model asymmetric ROC curves which occur when differences between TP and FP increase (or decrease) with increasing $c$. Continuous test outcomes can be analyzed using ordinal regression models by binning the data (Metz et al., 1998). Pepe proposed a generalized estimating equation approach for continuous data that allows specification of both $S_D(\cdot)$ and $\bar{S}_D(\cdot)$ (Pepe, 1997, 1998, 2000). Estimating equations for ROC models are based on observed and expected true positive and false positive rates. Models for clustered ordinal data can include non-independence working covariance structures, with variance estimates based on robust covariance estimation. Models for clustered continuous tests use an independence working correlation structure with variance estimates based on bootstrap resampling.

Bayesian models for ROC curve estimation of clustered data have been proposed for ordinal tests. Gatsonis describes hierarchical ordinal regression models when there is clustering within radiologists (Gatsonis, 1995). This approach extends the ordinal regression model by allowing random effects for accuracy and/or cutpoint parameters.

We propose a hierarchical Bayesian model for ROC curve estimation that uses random effects to account for clustering within individuals when test outcomes are continuous. Our approach specifies a parametric model for test outcomes and therefore requires attention to model fit. In return, the model provides information about the distribution of test outcomes. We embed a Truncated Normal (TN) model for SCL scores into a fully Bayesian hierarchical model and use random effects to account for patient-level correlation. We estimate our model using Markov chain Monte Carlo (MCMC) simulation.

3.1 The area under the ROC curve (AUC)

The overall accuracy of a test is commonly summarized by the area under the ROC curve (AUC). The AUC can be interpreted as the average true positive rate across the $(0, 1)$ range of false positive rates, or as the probability of correctly ranking a randomly chosen (diseased, not-diseased) pair. The AUC is 1 for a perfect test and 0.5 for a test that is no better than chance. The AUC based on the empirical ROC curve is asymptotically normally distributed, and standard errors are easily calculated (Bamber, 1975). When data are clustered, bootstrap estimation can be used to estimate the empirical AUC. When the ROC curve is estimated using ordinal regression with a probit link, the AUC is a simple function of regression coefficients and the AUC standard deviation can be estimated using Taylor series approximation (Hanley and McNeil, 1982). For Bayesian models that are fit using simulation methods, the posterior distribution of the AUC curve can be estimated using draws from the posterior distribution of model parameters.

3.2 Estimation of optimal cutpoints

Estimation of optimal cutpoints requires specification of optimal characteristics for the resulting dichotomous test. For example, the point on the ROC curve with $TP = (1 - FP)$ minimizes overall error. This cutpoint is appealing because it is objective; however, false negative and false positive results are rarely considered to have the same cost. Other optimal cutpoint definitions are necessarily subjective. Optimal
cutpoints can be selected based on the relative costs of false negative and false positive results (Halpern et al., 1996). Cutpoints that produce dichotomous tests with particular TP or FP rates might also be considered ‘optimal’. Because the goal of testing is predictive, cutpoints based on the diagnostic likelihood ratio (DLR) statistics may be more useful (Boyko, 1994). The DLR estimates the change in post-test odds of disease. For example, the diagnostic likelihood ratio positive (DLR+) is the multiplicative change in the post-test odds of being depressed after a positive test and is given by sensitivity/(1 − specificity). Similarly, the DLR− is the post-test odds of not being depressed after a negative test and is given by (1 − sensitivity)/specificity.

Optimal cutpoints are often estimated informally and presented without measures of their uncertainty. We know of no published methods for interval estimation of cutpoints. Given optimal properties, our method allows estimation of cutpoints with 95% probability intervals using simulated draws from the joint posterior distribution of model parameters.

4. Model

Suppose there are two disease states with D = 1 when diseased and D = 0 when not-diseased. We assume that the test result given disease status, Y | D, is approximately continuous with a unimodal distribution ranging from A to B. Let yit be the ith screening result for the ith subject with true disease state d_it; i = 1, . . . , N; t = 1, . . . , T. We use a TN distribution to model test outcomes because it allows the distribution of the scale outcome to be skewed in different directions for diseased and not-diseased groups and can accommodate symmetric or even relatively uniform distributions. If each patient contributes a single observation, then we can model yit given d_it as

\[ y_{it} | (D_{it} = d_{it}, \mu_{id_{it}}, \tau_{d_{it}}) \sim TN(\mu_{id_{it}}, 1/\tau_{d_{it}}), \quad A \leq y_{it} \leq B \]

\[ \mu_{id} = \alpha_d + \theta_d. \]

The parameters (\( \alpha_0 \), \( \tau_0 \), \( \theta_{10} \)) describe the test distribution when not diseased, and (\( \alpha_1 \), \( \tau_1 \), \( \theta_{11} \)) describe the test distribution when diseased. The random effects are linked across individuals via distributions for \( \theta_d \). The \( \theta \) parameters allow variability in the distribution of scale outcomes across individuals and correlation within individuals. We assume that the random effects \( \theta_d \) follow a bivariate normal distribution, which we implement using regression:

\[ \begin{pmatrix} \theta_{10} \\ \theta_{11} \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1/\xi_0 & \beta/\xi_0 \\ \beta/\xi_0 & \beta^2/\xi_0 + 1/\xi_1 \end{pmatrix} \right) \].

This exchangeable correlation structure within disease state implies that correlation between SCL scores is constant across time lags.

For an arbitrary cutpoint, c, an individual’s expected true positive rate, TP_i, is given by

\[ P(Y_i \geq c | D = 1) = 1 - \frac{\Phi((c - \alpha_1 - \theta_{11})\sqrt{\tau_1})}{\Phi((B - \alpha_1 - \theta_{11})\sqrt{\tau_1}) - \Phi((A - \alpha_1 - \theta_{11})\sqrt{\tau_1})}. \]

The expression for an individual’s false positive rate (FP_i) is similar and is based on \( \alpha_0 \), \( \theta_{10} \) and \( \tau_0 \). These individual rates are useful for model diagnostics. Our analytic results focus on the population-average TP.
and FP rates that are based on marginalizing individual rates across the random effects distribution. For example, the population-average TP rate is given by

\[
E(P(Y \geq c | D = 1)) = \int_{-\infty}^{+\infty} P(Y \geq c | D = 1) f(\theta_1) d\theta_1
\]

where \( f(\theta_1) = N(0, \beta^2/\xi_0 + 1/\xi_1) \).

The model is completed by specifying prior distributions for the remaining unknown parameters: \( \alpha_d \), \( \tau_d \), \( \xi_d \), and \( \beta \). For simplicity, we chose a Uniform prior for \( \beta \) over a wide range. Similarly, we chose a Uniform prior on \( \alpha_d \) with support over a range that is somewhat broader than the observable range of the data to allow for very skewed distributions. When using a TN, similar distributional shapes can be achieved from different combinations of \( \alpha_d \) and \( \tau_d \). Therefore, the range of \( \alpha_d \) should be restricted to ensure identifiability of \( \alpha_d \) and \( \tau_d \). We assume the precisions \( \tau_d \) and \( \xi_d \) are Gamma distributed with parameters \( a \) and \( b \). Finally, we assume that \( \alpha \), \( \tau \), \( \xi \), and \( \beta \) are mutually independent.

5. Analysis of RCT Data

We restricted our analysis of SCL accuracy to follow-up measures because at baseline patients had not yet initiated treatment. We therefore omitted data from 27 patients who had no follow-up information. Among the 509 patients with at least one follow-up, the percentage missing at each time point is: 1 month: 0.0%, 3 months: 7.1%, 6 months: 9.0%, 9 months: 12.4%, 12 months: 15.7%, 18 months: 20.2%, 24 months: 25.5%. We assume that study drop-out is independent of SCL accuracy. This assumption seems reasonable since we observed no association between the total number of follow-up observations and SCL accuracy at 1 month. The 1 month AUC statistic based on subjects with 2 or 3 observations is similar to AUC statistics based on individuals with 7 (AUC = 0.83) or 8 (AUC = 0.91) observations.

Analyses combine 2781 SCL scores from not-depressed patients and 322 SCL scores from depressed patients. Four patients were depressed at all follow-ups, 295 patients were not depressed at all follow-ups, and the remaining 210 patients had observations from both depression states. Figure 1 shows the empirical ROC curves at each time point. These curves demonstrate similar accuracy of SCL scores for detecting major depression across time points. We found no evidence of trends in SCL accuracy over time. The empirical AUC statistics for each time point (with 95% confidence interval) are: 1 month: 0.90(0.86, 0.94), 3 months: 0.86(0.81, 0.91), 6 months: 0.89(0.85, 0.93), 9 months: 0.91(0.88, 0.95), 12 months: 0.91(0.88, 0.94), 18 months: 0.91(0.85, 0.96), 24 months: 0.95(0.92, 0.97). The bootstrap estimate of the overall AUC statistics is 0.90 with 95% confidence interval (0.89, 0.92).

We observed moderately strong within-patient correlation of SCL scores over time (Table 2). We also examined serial correlation by depression status. There was no evidence of serial correlation when depression was present. When depression was absent, correlation was only slightly higher at shorter lags (near 0.40) than for later lags (near 0.35). Therefore, an exchangeable correlation structure provides a reasonable model for these data.

We estimated five different cutpoints on the SCL scale. The first balances misclassification errors so that sensitivity equals specificity. The second produces a binary test that is 80% sensitive, and the third produces a binary test that is 80% specific. The fourth and fifth are based on positive and negative diagnostic likelihood ratio statistics. The fourth cutpoint corresponds to a DLR$^+$ = 2, or a 200% increase in the odds of depression following a positive test. The fifth cutpoint corresponds to a DLR$^-$ = 0.5 or a 50% decrease in the odds of depression following a negative test.
Table 2. Correlation of consecutive SCL measures

<table>
<thead>
<tr>
<th>Transition</th>
<th>Number of subjects</th>
<th>Number of observations</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depressed at both times</td>
<td>38</td>
<td>56</td>
<td>0.39</td>
</tr>
<tr>
<td>Depressed to not depressed</td>
<td>189</td>
<td>217</td>
<td>0.35</td>
</tr>
<tr>
<td>Not depressed to depressed</td>
<td>163</td>
<td>190</td>
<td>0.36</td>
</tr>
<tr>
<td>Not depressed at both times</td>
<td>477</td>
<td>2054</td>
<td>0.47</td>
</tr>
</tbody>
</table>

Table 3. Estimated model parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mode</th>
<th>95% HPD*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not depressed normal mean</td>
<td>(\alpha_0)</td>
<td>-0.86</td>
</tr>
<tr>
<td>SD of (y_{id=0})</td>
<td>1/(\tau_0)</td>
<td>0.74</td>
</tr>
<tr>
<td>SD of (\theta_{id=0})</td>
<td>1/(\xi_0)</td>
<td>1.58</td>
</tr>
<tr>
<td>Depressed normal mean</td>
<td>(\alpha_1)</td>
<td>1.51</td>
</tr>
<tr>
<td>SD of (y_{id=1})</td>
<td>1/(\tau_1)</td>
<td>0.58</td>
</tr>
<tr>
<td>SD of (\theta_{id=1})</td>
<td>(\beta^2/\xi_0 + 1/\xi_1)</td>
<td>0.60</td>
</tr>
<tr>
<td>Correlation (\text{Corr}(\theta_{id=0}, \theta_{id=1}))</td>
<td>(\beta(\beta^2 + \xi_0/\xi_1)^{-1/2})</td>
<td>0.61</td>
</tr>
</tbody>
</table>

*Highest posterior density region

5.1 Estimation and diagnostics

We used a Uniform\([-1, 5]\) prior for \(\alpha_d\), which allows very skewed distributions over the [0, 4] range of SCL scores. Because we lacked information on the range of \(\beta\), we assumed \(\beta \sim \text{Uniform}[-100, 100]\). We assumed that the precision parameters \((\tau_d, \xi_d)\) were distributed Gamma(0.001, 0.001). Parameters were estimated using MCMC estimation (Gelman et al., 1995; Gilks et al., 1996) (see the Appendix for details).

We assessed convergence using several criteria (see Appendix), including trace plots from multiple starting points. We used Q–Q plots to compare the observed distribution of SCL scores to estimated distributions.

5.2 Estimated densities and ROC curves

Estimation of the SCL densities, and therefore ROC curves, requires incorporation of both the within-subject variability \((1/\tau_d)\) and between-subject variability \((1/\xi_0 \text{ and } \beta^2/\xi_0 + 1/\xi_1)\). Study-specific (SS) densities and ROC curves were estimated directly using MCMC draws from the parameter space, including each individual’s estimated random effects. Population-average densities and ROC curves were estimated by marginalizing over the random effects distribution, via Monte Carlo integration. Population-average estimates are more generalizable than study-specific estimates because they incorporate expected population variability. We do not calculate subject-specific estimates, because they do not incorporate random effects and therefore underestimate SCL variability and overestimate SCL accuracy.

5.3 Results

Q–Q plots, based on SS estimates, demonstrated very good model fit. The model fit equally well across time points. Estimated model parameters are given in Table 3.

Figure 2 shows estimated population-average (PA) SCL distributions, demonstrating right-skewness in the not-depressed group and approximate symmetry in the depressed group. As shown in Figure 3,
the PA ROC curve is attenuated relative to the SS ROC curve. This attenuation is also reflected in AUC statistics. The SS AUC is 0.90 with 95% credible interval (0.88, 0.92), nearly identical to the bootstrap AUC estimate. The PA AUC is 0.86 with 95% credible interval (0.83, 0.88). Overall, these figures show good separation of SCL distributions and we conclude that the SCL is an accurate self-administered screening tool for DSM-IV major depression.

Table 4 shows estimated clinical cutpoints for identifying depressed and not-depressed patients based on the PA ROC curve. The estimated $Q^*$ statistic, the point at which sensitivity is equal to specificity, is 0.78 (95% HPD: (0.75, 0.81). The estimated cutpoint to achieve a $DLR^- = 0.5$ was 1.38. Estimation of the cutpoint for $DLR^+ = 2$ requires additional specification, because for estimated ROC curves there
Table 4. Optimal SCL cutpoints

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Estimated SCL cutpoint</th>
<th>95% HPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>80% sensitivity</td>
<td>0.77</td>
<td>(0.62, 0.94)</td>
</tr>
<tr>
<td>80% specificity</td>
<td>0.92</td>
<td>(0.80, 1.06)</td>
</tr>
<tr>
<td>Sensitivity = specificity</td>
<td>0.85</td>
<td>(0.74, 0.96)</td>
</tr>
<tr>
<td>Q* statistic</td>
<td>0.78</td>
<td>(0.75, 0.81)</td>
</tr>
<tr>
<td>$DLR^{-} = 0.5$</td>
<td>1.38</td>
<td>(1.22, 1.54)</td>
</tr>
<tr>
<td>$DLR^{+} = 2$, highest sensitivity</td>
<td>0.31</td>
<td>(0.25, 0.39)</td>
</tr>
<tr>
<td>$DLR^{+} = 2$, highest specificity</td>
<td>3.66</td>
<td>(0.26, 3.90)</td>
</tr>
</tbody>
</table>

were often two possible cutpoints, one associated with high sensitivity and another associated with high specificity. These results show that when dichotomizing SCL results, the motivation behind testing should drive the choice of the cutpoint that determines a positive SCL score.

6. Discussion

The proposed method allowed us to estimate the accuracy of SCL as a screen for residual depression using longitudinal patient data. We were also able to estimate useful cutpoints for the SCL scale with interval estimates indicating their precision. The Truncated Normal (TN) model fits the data very well, accommodating both the skewed distribution for not-depressed patients and the symmetric distribution for depressed patients. We used simulation-based approaches for estimation of the posterior distribution of model parameters, including population average effects of model parameters based on integration over the random effects distributions. Simulation was also required for quantile estimation that was used for both model diagnostics and estimation of ROC curves, cutpoints and the area under estimated ROC curves.

Currently, each of these steps requires specialized software.

Realistic summary measures should fully account for expected population variability. Population average estimates incorporate between-subject variability and thereby provide generalizable estimates of model accuracy. Study-specified (SS) estimates condition on the random effects for the observed sample and only account for observed sample variability. Our population average estimates of SCL accuracy were attenuated relative to SS estimates. This attenuation reflects the relatively wide variability in SCL scores among the not-depressed group. Fully incorporating this variability into estimates resulted in greater overlap between SCL distributions for depressed and not-depressed groups. SS estimates are useful for examining model fit, but may overestimate test accuracy. We did not examine subject-specific ROC curves because these ignore between individual variability and therefore overestimate test accuracy.

Results from the random effects model provided insight into the distribution of SCL scores over time. We found high correlation in SCL scores over time. That is, some patients tended to endorse symptom questions and have high SCL scores regardless of their depression state. Clinically, this suggests that screening for depression could be improved over time by using changes from previous screening outcomes.

Several aspects of our model are specific to the data we analyze. The model we develop assumes that there are individuals who are observed in both depressed and not-depressed states. If depression state were constant over time within individuals, then our model would need to be simplified to include only one random effect per individual, with the variance of this random effect allowed to depend on depression state. We also make two simplifying assumptions that might not hold for other data analyses. First, we assumed an exchangeable correlation structure. Because inappropriate models for correlation can affect interval estimates, correlation structures must be carefully selected and validated. Second, we assume
that study dropout is unrelated to SCL accuracy. Estimated ROC curves may be biased if study dropout
is non-random. For example, systematic dropout of depressed individuals with high SCL scores could
attenuate estimated ROC curves, while systematic dropout of not-depressed individuals with high SCL
scores, though unlikely, could inflate ROC estimates. We used simple methods to check these assumptions
and found that they were reasonable for our data. Extensions to allow more general covariance structures
and to adjust for differential dropout would be relatively complex.

Other extensions to our model may be useful. Including covariates in the model would allow
exploration of important issues of differential test accuracy and estimation of different optimal cutpoints
for subgroups of patients. Covariate effects can be included by extending the regression model for the
mean of the normal distribution that is truncated. Because shifts in the normal mean alone affect the shape
of the TN distribution, the variance may also need to be modeled as a function of covariates.

Finally, it would be useful to extend the model to account for errors in the gold standard via latent class
analysis (e.g. Formann and Kohlmann, 1996). This approach would treat items of the SCID as a set of
imperfect measures of depression, with true depression status treated as an underlying latent variable with
two classes (major depression, yes or no). The latent class modeling approach would allow for exploration
of the effects of misclassification of the SCID for diagnosing depression. This is especially important for
assessing the accuracy of psychological screening scales, since clinical diagnosis inherently depends on
patient recall and generally cannot be made based on ‘objective’ laboratory tests.

ACKNOWLEDGEMENTS

This work was supported by research grants from the NIMH (RO1 MH 51338 and RO1 MH 41739).
The authors would also like to thank Greg Simon, MD and the Liaison Psychiatry Team for their support of
this work, and reviewers for improving the quality of this manuscript through their constructive comments.

APPENDIX

The random effects precisions, \( \xi_d \), were drawn directly from their full conditional (Gamma)
distributions. Draws for the full conditional distributions of remaining parameters were simulated using
Metropolis (random walk) steps (Gelman et al., 1995). We started the MCMC samplers at five different
points in the parameter space. Three mean-pairs starting values were used with initial correlation set to
zero: (1) \( \mu_0^{(0)} = 0, \mu_1^{(0)} = 2 \); (2) \( \mu_0^{(0)} = \mu_1^{(0)} = 2 \) and (3) \( \mu_0^{(0)} = 2, \mu_1^{(0)} = 0.5 \). Two additional chains
were started at \( \mu_0^{(0)} = \mu_1^{(0)} = 2 \) with initial correlation set to \( \rho^{(0)} = -0.75 \) and \( \rho^{(0)} = 0.75 \). Starting
values for the variance parameters were set larger than expected for each chain: \( \tau_0 = \tau_1 = \xi_0 = \xi_1 = 2 \).
Correlated starting values for random effects were drawn from the prior distribution with \( \xi_0 = \xi_1 = 2 \)
and \( \rho = \rho^{(0)} \). Each chain was run for 10000 iterations. We allowed 1000 iterations for burn-in, and kept
every 10th draw from the remaining 9000 for final analyses. Final estimates were based on 4500 simulated
draws from five independent samplers that used different starting points.

We used convergence diagnostics provided by CODA (Best et al., 1995), including the Gelman and
Rubin statistic (Gelman and Rubin, 1992) for multiple chains.

Posterior modes and highest posterior density intervals (HPDs) were estimated from simulated draws
from the full posterior distribution using density smoothing via SAS PROC KDE (SAS Institute, 2000).

ROC curves are based on the probability of being above a particular cutpoint in depressed versus not-
depressed populations. We estimated ROC curves at each draw by estimating the true positive rate for
false positive rates in \([0.01, 0.02, 0.03, \ldots, 0.97, 0.98, 0.99]\). The final estimated ROC curves are based
on the modal true positive rates across the fixed set of false positive rates.
REFERENCES


[Received July 20, 2001; first revision December 17, 2001; second revision January 23, 2002; accepted for publication February 1, 2002]