Original Paper

# Maximizing the Value of Mobile Health Monitoring by Avoiding Redundant Patient Reports: Prediction of Depression-Related Symptoms and Adherence Problems in Automated Health Assessment Services

John D Piette[1], PhD, ScM; Jeremy B Sussman[1], MD; Paul N Pfeiffer[2], MD; Maria J Silveira[1], MD; Satinder Singh[3], PhD; Mariel S Lavieri[4], PhD

[1]VA Center for Clinical Management Research and Division of General Medicine, Department of Internal Medicine, University of Michigan, Ann Arbor, MI, United States

[2]VA Center for Clinical Management Research and Department of Psychiatry, Ann Arbor VA Healthcare System and University of Michigan, Ann Arbor, MI, United States

[3]Artificial Intelligence Laboratory, Department of Electrical Engineering and Computer Science, College of Engineering, University of Michigan, Ann Arbor, MI, United States

[4]Deparment of Industrial and Operations Engineering, College of Engineering, University of Michigan, Ann Arbor, MI, United States

**Corresponding Author:**
John D Piette, PhD, ScM
VA Center for Clinical Management Research and Division of General Medicine
Department of Internal Medicine
University of Michigan
PO Box 130170
Ann Arbor, MI, 48113-0170
United States
Phone: 1 734 936 4787
Fax: 1 734 936 8944
Email: jpiette@umich.edu

## Abstract

**Background:** Interactive voice response (IVR) calls enhance health systems' ability to identify health risk factors, thereby enabling targeted clinical follow-up. However, redundant assessments may increase patient dropout and represent a lost opportunity to collect more clinically useful data.

**Objective:** We determined the extent to which previous IVR assessments predicted subsequent responses among patients with depression diagnoses, potentially obviating the need to repeatedly collect the same information. We also evaluated whether frequent (ie, weekly) IVR assessment attempts were significantly more predictive of patients' subsequent reports than information collected biweekly or monthly.

**Methods:** Using data from 1050 IVR assessments for 208 patients with depression diagnoses, we examined the predictability of four IVR-reported outcomes: moderate/severe depressive symptoms (score ≥10 on the PHQ-9), fair/poor general health, poor antidepressant adherence, and days in bed due to poor mental health. We used logistic models with training and test samples to predict patients' IVR responses based on their five most recent weekly, biweekly, and monthly assessment attempts. The marginal benefit of more frequent assessments was evaluated based on Receiver Operator Characteristic (ROC) curves and statistical comparisons of the area under the curves (AUC).

**Results:** Patients' reports about their depressive symptoms and perceived health status were highly predictable based on prior assessment responses. For models predicting moderate/severe depression, the AUC was 0.91 (95% CI 0.89-0.93) when assuming weekly assessment attempts and only slightly less when assuming biweekly assessments (AUC: 0.89; CI 0.87-0.91) or monthly attempts (AUC: 0.89; CI 0.86-0.91). The AUC for models predicting reports of fair/poor health status was similar when weekly assessments were compared with those occurring biweekly (*P* value for the difference=.11) or monthly (*P*=.81). Reports of medication adherence problems and days in bed were somewhat less predictable but also showed small differences between assessments attempted weekly, biweekly, and monthly.

**Conclusions:** The technical feasibility of gathering high frequency health data via IVR may in some instances exceed the clinical benefit of doing so. Predictive analytics could make data gathering more efficient with negligible loss in effectiveness. In particular, weekly or biweekly depressive symptom reports may provide little marginal information regarding how the person is doing relative to collecting that information monthly. The next generation of automated health assessment services should use data mining techniques to avoid redundant assessments and should gather data at the frequency that maximizes the value of the information collected.

***(J Med Internet Res 2013;15(7):e118)*** doi: 10.2196/jmir.2582

**KEYWORDS**

cellular phone; telemedicine; depression; self-care

## Introduction

Clinicians and health care payers increasingly look to mobile health services such as Interactive Voice Response (IVR) as tools for monitoring patients' status between face-to-face encounters and identifying individuals who need attention to prevent acute events [1-3]. Multiple studies have shown that IVR monitoring yields actionable and reliable clinical information even on sensitive topics such as mental health and substance abuse [4-11]. Moreover, patients are willing to complete regular IVR assessments over extended periods of time, even when challenged by chronic illness, age, poverty, low literacy, and psychiatric problems [12,13].

While IVR has significant potential to increase the information base of proactive care management, the design of automated monitoring services can have negative consequences that should be carefully considered when deciding the frequency and content of each assessment call. Studies suggest that patients may tire of frequent IVR assessments [12-15], particularly if they are asked repeatedly for information about health or self-care problems that have not changed. At the same time, many patients have a large number of health problems associated with multiple chronic conditions [16,17]. For such patients, current alternatives to the typical disease-specific focus include substantially increasing the length of each assessment, increasing the frequency of assessment calls, focusing on a broader number of problems but with less depth on each, or focusing only on cross-cutting issues such as medication adherence or physical activity. Each of these strategies introduces new challenges to sustaining patient engagement or the quality of information for clinical decisions. As with other types of patient contact [18-21], the timing and content of IVR monitoring is almost always based on expert opinion and static flow diagrams. As such, these systems have not achieved their full potential as a strategy for cost-effectively increasing patients' access to between-visit monitoring and self-care support.

While frequent (eg, weekly or daily) IVR assessment calls may be necessary to detect fluctuations in important health indicators, what if a patient's IVR assessment reports could be predicted based on the information that he or she provided in prior calls? For example, if a patient has consistently reported perfect medication adherence over multiple prior IVR assessments, what would be the probability that they would report something different today? Data mining is a set of analytic techniques designed to extract latent information from data in order to make predictions about the future [22,23]. In the context of IVR, data mining could help identify when patients' answers are so stable that the same questions are not worth asking again, or when there are changes in the patient's status indicating the need for more intensive probing. Using information about such patterns, adaptive mobile health monitoring programs could be developed that automatically adjust the frequency and content of assessments so that they provide the most useful information for guiding patient counseling and clinical follow-up.

We used one approach to data mining in order to examine data from 1050 IVR assessments of 208 patients with depression diagnoses. All patients received IVR calls at regular intervals, during which they completed the Patient Health Questionnaire (PHQ-9) [24,25], a widely used and validated depression assessment scale. Also, patients repeatedly answered questions regarding their antidepressant medication adherence, perceived general health, and days in bed due mental health problems. Given the large number of serial reports from each patient, we examined the predictability of patients' IVR responses. Specifically, for each patient we identified the five most recent weekly, biweekly, and monthly assessments. We used those data plus other information collected during prior assessments and at the time of the patient's enrollment to determine the extent to which health reports were predictable and whether that predictability varied according to the frequency of attempted assessment calls. Based on these analyses, we determined whether less frequently collected data (eg, biweekly or monthly) provided as much information about patients' status as information collected weekly, thereby making it possible to decrease the frequency of IVR calls or to change their focus to other important health indicators. More generally, we sought to determine whether data mining techniques might inform automated assessments that repeatedly measure patients' health status, so that the most clinically useful, nonredundant information is collected.

## Methods

### Patient Eligibility and Recruitment

Patients were enrolled between March 2010 and January 2012 from 13 university-affiliated and community-based primary care practices. To be eligible, patients had to have two primary care visits in the previous 2 years, at least one in the previous 13 months, and either a depression diagnosis listed in clinical records or an antidepressant prescription plus a diagnosis of depression listed in billing data. Patients with schizophrenia, psychosis, delusional disorder, bipolar disorder, or dementia were excluded. Potential participants were mailed an introductory letter that was followed by a screening and recruitment telephone call. Patients who provided informed consent were enrolled in the IVR system and mailed additional program information, including materials describing effective communication with informal caregivers and clinicians. The study was approved by the human subjects committees of the University of Michigan and Ann Arbor VA Healthcare System. More information about the intervention and patients' engagement in the IVR calls has been published elsewhere [13].

## IVR Monitoring Protocol

Detailed information about the IVR call content and functioning are available by contacting the authors. In brief, each week that an assessment was scheduled, the system made up to three attempts to contact the patient on up to three different patient-selected day/time combinations. The content of the calls was developed with input from psychiatrists, primary care providers, and experts in IVR program design and health behavior change. Every call included an assessment of patients' depression symptoms using the PHQ-9 [24]. The PHQ-9 is a 9-item questionnaire that is sensitive and specific with respect to other established measures of major depression. Scores are associated with physical functioning, sick days, and health care use [24]. Because self-rated health status is correlated with patients' service use and mortality risk [26-28], they were asked the standard item, "Thinking about your overall health, how were you feeling this past week (excellent, very good, good, fair, poor)?" Medication adherence was assessed by asking: "How often during the past week did you take your depression medication exactly as prescribed (always, most of the time, less than half of the time, rarely or never)?" Finally, during each assessment, patients were asked: "This past week, did you ever stay in bed all or most of the day because of your mental health (yes versus no)?" Calls used tree-structured algorithms to present recorded queries and tailored information that was invariant across patients and over time. Based on patients' responses, they received tailored advice for managing their self-care. For example, patients' received messages tailored according to their recent trajectories in depression scores (trending positive, negative, or stable and by how much), including messages such as the following:

> *It sounds like you're still experiencing some serious symptoms of depression. Remember that if you're prescribed a medication for depression, it's important that you keep taking it exactly as prescribed to keep your depression from getting worse. Sometimes it takes awhile for a depression medication to work, so if you have been on your current medication for less than 8 weeks, try to be patient and see if you start to see some improvement. If you've been on the same medication for more than 8 weeks and you're still not feeling okay, your doctor wants to know. You should make an appointment with your doctor to talk about some other treatment options. I'll give you the phone number of your doctor's office at the end of this call.*

Clinicians received fax alerts identifying patients reporting health problems requiring follow-up before their next outpatient encounter. For patients enrolling with a family caregiver, those caregivers received automatic updates by IVR and email with suggestions regarding how they could support the patient's self-management.

## Outcomes of Interest

For each assessment, we created binary indicators for each of the four outcomes reported: (1) moderate/severe depressive symptoms indicated by a PHQ-9 score of ≥10; (2) fair or poor perceived general health status; (3) poor antidepressant adherence, ie, rarely or never taking antidepressant medication as prescribed; and (4) spending days in bed in the past week due to mental health problems.

## Analytic Sample Definition and Analyses

In order to determine the predictability of patients' assessment reports based on the content and frequency of prior assessments, we identified the subset of patients with one or more "index" assessments meeting the following criteria: (a) five completed prior assessments immediately preceding the index assessment and collected with the program's normal frequency of weekly assessment attempts; (b) five completed prior assessments with a 2-week minimum gap between each one; and (c) five completed prior assessments with a minimum 4-week gap between each one. A total of 1050 index assessments for 208 unique patients were identified.

In addition to linking each index assessment to prior assessment information, index assessments also were linked with information about that patient's sociodemographic and clinical characteristics collected at the time of program initiation. Those baseline data included patients' age, gender, educational attainment, baseline depressive symptom severity score (ie, measured using the PHQ-9 minus the item asking about suicidal ideation [29]), self-reported hospital admission in the year prior to program entry, physical functioning as measured by the SF-12 [30], and the number of comorbid chronic medication conditions.

In initial analyses, we examined the correlation across the four health indicators reported within each index assessment, and we calculated the alpha reliability of patients' IVR-reported PHQ-9 scores. We then examined the proportion of patients reporting each health problem in the index assessment when the same problem was reported in the one or in both of the most recent prior assessments assuming weekly, biweekly, or monthly assessment attempts. For example, we examined measures of association between patient reports of moderate/severe depressive symptoms (PHQ-9 ≥10) and similarly high PHQ-9 scores in the most recent assessment or both of the two most recent assessments (assuming weekly, biweekly, and monthly assessment calls).

Finally, we fit multivariate logistic regression models predicting each of the four health indicators as reported in index assessments. Each model included patients' baseline sociodemographic and clinical characteristics as defined above, as well as information about that same health indicator and the other three health indicators reported in five prior assessments collected assuming a periodicity of weekly, biweekly, or monthly call attempts. Serial indicators designed to capture additional information about trends in patients' depression scores (eg, the number of weeks since program entry and prior number of completed assessments) also were considered as potential predictors. For models predicting moderate/severe depressive symptoms, fair/poor health, and days in bed, these additional variables had no discernible marginal predictive value in the context of the multiple prior, ordered indicators of the patient's health and self-care. However, an indicator for weeks since program entry was a marginally significant predictor of patients' medication adherence and was retained in the models used as the basis of ROC curves predicting patient reports of poor antidepressant medication adherence.

When fitting each of the three models, we used two strategies to prevent overfitting to the current dataset. First, we used 10-fold cross validation, in which the model was fit 10 times based on random 90% training samples and then used to predict the outcomes in mutually exclusive 10% test samples. Second, for each of the ten replications, we used stepwise regression (with a *P* value of .20 for removal) to identify the most significant subset of candidate predictors. All models also adjusted for clustering of assessment responses by patient.

The predictive significance of the three models for each outcome was compared graphically to one another and to a model with only baseline information using Receiver Operator Characteristic (ROC) curves. We also compared the area under the curve (AUC) across ROCs and calculated each AUC's 95% confidence interval [31]. To illustrate the potential predictive accuracy of the best model for each outcome, we report the sensitivity and specificity at the point on the ROC curve with the highest proportion of outcomes correctly predicted.

# Results

## Patient Characteristics

Patients were on average 52.2 years of age. Most were women, white, and married (Table 1). Patients reported a mean of 2.4 comorbid chronic conditions including hypertension (50.0%), arthritis (49.5%), chronic lung disease (33.2%) and back pain (42.1%). Roughly a third (33.2%) of patients had moderate or severe depressive symptoms at baseline; those patients were somewhat younger on average at the time of program enrollment than patients with mild depressive symptoms.

## Co-Occurrence of Reported Health Problems Within IVR Assessments

Patients reporting a given problem during their IVR assessments were more likely to report other concurrent problems as well. For example, compared to patients reporting mild depressive symptoms, those reporting moderate/severe depressive symptoms were more likely also to report staying in bed all or most of the day due to mental health problems (27% versus 8%) and that their general health was either fair or poor (47% versus 14%, both *P*<.001 after adjusting for clustering by patient). Similarly, patients reporting being bedbound during the past week due to mental health problems were significantly more likely than other patients to rate their health as fair or poor during the same assessment (29% versus 20%, *P*<.001). Patients reporting that they rarely or never took their medication as prescribed were more likely than other patients to report poor general health (28% versus 17%; *P*<.001).

## Bivariate Relationship Between IVR Reports and Prior Reports of the Same Outcome

The internal reliability of the PHQ-9 was excellent (alpha=.87). Patients were substantially more likely to report moderate/severe depressive symptoms if they reported similar information in prior assessments (Table 2). For example, while patients reported moderate/severe depressive symptoms in 21.5% of all assessments, they did so 70.3% of the time when they also reported similarly high symptoms on their most recent assessment, and 83.3% of the time when they reported moderate/severe depressive symptoms during both of their most recent assessments, assuming weekly assessment attempts. Ninety-one percent of patients whose most recent weekly PHQ-9 score was <10 also had a score <10 on their index assessment. Assuming weekly assessment attempts, a similar pattern was observed with respect to the autocorrelation of patients' reported general health status, medication adherence, and days in bed due to mental health problems.

In general, assessments collected biweekly or monthly were only somewhat less correlated with subsequent reports than information collected assuming weekly assessment attempts. For example, 58.8% of index assessments in which the patient reported moderate/severe depressive symptoms had similarly high levels in the two most recent assessments collected assuming weekly attempts, as compared to 53.4% on the two prior assessments collected biweekly, and 51% on the two prior assessments collected monthly.

## Predictive Models

### *Moderate/Severe Depression*

ROC curves for models predicting patients' depressive symptoms were highly predictive with an AUC≥0.89 regardless of whether prior assessments were attempted weekly, biweekly, or monthly (Figure 1 and Table 3). In Figure 1, the blue line represents weekly assessment attempts, the green line represents biweekly attempts, and the red line represents monthly attempts. The yellow line represents the ROC curve for the model predicting depressive symptoms using baseline data only. All other models also included baseline clinical and sociodemographic information. While the AUC for weekly assessments was significantly different than either biweekly (*P*<.001) or monthly assessments (*P*<.001), there was no statistically significant difference in the AUC for biweekly compared to monthly calls (*P*=.36).

The AUC for the model assuming weekly assessment attempts was .91 (95% CI 0.89, 0.93). At the point on the ROC curve with the greatest number of reports correctly classified (ie, a probability of moderate/severe depression=.50), 88.4% of assessments were classified correctly with a sensitivity of .68 and a specificity of .94. As expected, regardless of the frequency of assessment attempts, patients' prior PHQ-9 scores were the strongest predictor of index assessment scores ≥10, although prior IVR reports regarding general health status, baseline depressive symptom severity, baseline physical functioning, and the number of comorbidities reported at baseline also were significant independent predictors of patients' depression status.

### General Health Status

Similar to patients' reports of their depressive symptoms, reports of perceived general health status were highly predictable based on prior information (Figure 2). In Figure 2, the blue line represents weekly assessment attempts, the green line represents biweekly attempts, and the red line represents monthly assessment attempts. The yellow line represents the prediction based on baseline data only. All other models also included baseline clinical and sociodemographic information.

The AUC for the model assuming weekly assessment attempts was 0.88 (95% CI 0.86, 0.91). The AUC for that model was not statistically different from the one assuming biweekly attempts (*P*=.11) or assessments collected monthly (*P*=.81). Prior reports of perceived health status were the strongest predictors, although prior information about days in bed due to mental health problems and about medication adherence problems also were consistently retained in logistic models as predictors of patients' index assessment reports of fair/poor health. With respect to the model assuming weekly assessment attempts, the cutoff indicating a probability of fair/poor health=.50 correctly classified 87% of all index assessments, with a sensitivity of .58 and a specificity of .95.

**Table 1.** Patient characteristics (cell entries, aside from N, are either column percent or mean [SD]).

| | Depressive symptom severity[a] | | | |
|---|---|---|---|---|
| | Total | Moderate/Severe | Mild | *P* value |
| N | 208 | 69 | 139 | |
| Age in years | 52.2 (12.5) | 50.6 (12.0) | 53.7 (12.8) | .04 |
| Female | 79.0 | 77.9 | 80.0 | .72 |
| White | 90.0 | 89.5 | 90.5 | .81 |
| Married | 60.0 | 57.9 | 62.1 | .55 |
| More than high school | 79.5 | 75.8 | 83.2 | .21 |
| Prior hospitalization[b] | 21.6 | 24.2 | 19.0 | .38 |
| Number of diagnoses | 2.4 (1.7) | 2.6 (1.7) | 2.2 (1.6) | .09 |
| Hypertension | 50.0 | 55.8 | 44.2 | .11 |
| Cardiovascular disease | 8.4 | 10.5 | 6.3 | .30 |
| Stroke | 4.2 | 4.2 | 4.2 | 1.00 |
| Arthritis | 49.5 | 52.6 | 46.3 | .38 |
| Chronic lung disease | 33.2 | 41.1 | 25.3 | .02 |
| Back pain | 42.1 | 43.2 | 41.1 | .77 |
| Physical functioning[c] | 39.6 (13.8) | 37.8 (14.2) | 41.4 (13.3) | .07 |

[a] PHQ-9: 9-item Patient Health Questionnaire score ≥10 or <10.

[b] 1+ hospitalizations in the year prior to enrollment.

[c] Physical Functioning: 12-item Medical Outcome Study Short Form Physical Composite Summary. Scores range from 0 to 100 with higher scores indicating greater functioning.

**Table 2.** Variation in problem reports by the number and frequency of prior reports of the same problem.

| | Weekly | | Biweekly | | Monthly | |
|---|---|---|---|---|---|---|
| | 1 report[a] | 2 reports[b] | 1 report | 2 reports | 1 report | 2 reports |
| **Moderate/Severe Depression** [c] | | | | | | |
| % with Report[d] | 21.5 | 21.5 | 21.5 | 21.5 | 21.5 | 21.5 |
| Sensitivity[e] | 69.6 | 58.8 | 69.1 | 53.4 | 64.2 | 51.0 |
| Specificity[f] | 92.0 | 96.8 | 90.5 | 95.4 | 89.8 | 95.8 |
| PPV[g] | 70.3 | 83.3 | 66.5 | 76.2 | 63.3 | 77.0 |
| NPV[h] | 91.7 | 89.6 | 91.5 | 88.2 | 90.2 | 87.7 |
| **Fair/Poor Health** | | | | | | |
| % with Report | 21.4 | 21.4 | 21.4 | 21.4 | 21.4 | 21.4 |
| Sensitivity | 67.2 | 57.4 | 67.2 | 57.4 | 67.7 | 55.9 |
| Specificity | 90.5 | 96.7 | 89.4 | 96.7 | 89.7 | 96.4 |
| PPV | 65.9 | 82.4 | 63.4 | 82.4 | 64.2 | 80.9 |
| NPV | 91.0 | 89.3 | 90.9 | 89.3 | 91.0 | 88.9 |
| **Poor Adherence** | | | | | | |
| % with Report | 18.6 | 18.6 | 18.6 | 18.6 | 18.6 | 18.6 |
| Sensitivity | 55.5 | 43.2 | 54.2 | 42.6 | 58.4 | 39.6 |
| Specificity | 90.4 | 96.2 | 91.2 | 96.6 | 90.1 | 96.5 |
| PPV | 57.0 | 72.0 | 58.3 | 74.2 | 57.3 | 71.8 |
| NPV | 89.9 | 88.1 | 89.7 | 88.1 | 90.5 | 87.5 |
| **In Bed Due to Mental Health** | | | | | | |
| % with Report | 12.9 | 12.9 | 12.9 | 12.9 | 12.9 | 12.9 |
| Sensitivity | 45.4 | 24.2 | 39.5 | 17.5 | 30.3 | 11.7 |
| Specificity | 91.8 | 97.3 | 91.0 | 97.3 | 90.7 | 97.4 |
| PPV | 45.0 | 55.8 | 39.5 | 47.7 | 32.7 | 38.9 |
| NPV | 91.9 | 90.0 | 91.0 | 89.2 | 89.7 | 88.6 |

[a] Patient also reported the same health problem in the most recent assessment during the time frame.

[b] Patient also reported the same health problem in the two most recent assessments during the time frame.

[c] PHQ-9 score ≥10.

[d] Percentage of all index assessments in which that health problem was reported.

[e] Proportion of index assessments reporting that health problem that also had the problem reported in the prior assessment(s).

[f] Proportion of index assessments not reporting that health problem that also were negative in the prior assessment(s).

[g] PPV: Positive Predictive Value; given that the problem was reported in the prior assessment(s), the proportion reporting that problem in the index assessment.

[h] NPV: Negative Predictive Value; given that the problem was not reported in the prior assessment(s), the proportion of index assessments that also did not report the problem.

**Table 3.** Area under the Receiver Operator Characteristic (ROC) curve for logistic models predicting each health indicator assuming different assessment frequencies.

| | AUC[a] | 95% CI |
|---|---|---|
| **Moderate/Severe Depression** [b] | | |
| Weekly | 0.9139 | 0.8931, 0.9348 |
| Biweekly | 0.8887 | 0.8655, 0.9119 |
| Monthly | 0.8873 | 0.8630, 0.9116 |
| Baseline data only | 0.7396 | 0.7010, 0.7782 |
| **Fair/Poor General Health** | | |
| Weekly | 0.8840 | 0.8581, 0.9100 |
| Biweekly | 0.8758 | 0.8477, 0.9039 |
| Monthly | 0.8822 | 0.8543, 0.9101 |
| Baseline data only | 0.6760 | 0.6367, 0.7154 |
| **Poor Antidepressant Adherence** | | |
| Weekly | 0.8396 | 0.8035, 0.8757 |
| Biweekly | 0.8268 | 0.7899, 0.8637 |
| Monthly | 0.8350 | 0.8000, 0.8701 |
| Baseline data only | 0.7578 | 0.7162, 0.7993 |
| **In Bed Due to Mental Health** | | |
| Weekly | 0.7522 | 0.7058, 0.7986 |
| Biweekly | 0.6872 | 0.6358, 0.7385 |
| Monthly | 0.7197 | 0.6716, 0.7677 |
| Baseline data only | 0.6029 | 0.5542, 0.6515 |

[a] Area Under the Curve.

[b] PHQ-9 score ≥10.

### Poor Antidepressant Adherence

While the overall predictive power was somewhat lower across models predicting reports of medication adherence problems, those models also showed that information collected biweekly or monthly was similar in its correlation with index assessment reports compared to information collected weekly (Table 3 and Figure 3). In Figure 3, the blue line represents weekly assessment attempts, the green line represents biweekly attempts, and the red line represents monthly attempts. The yellow line represents the ROC curve for the model predicting poor adherence using baseline data only. All other models also included baseline clinical and sociodemographic information.

The AUC for the model based on weekly assessments was 0.84 (95% CI 0.80, 0.88). The AUC for that model was not significantly different compared to either biweekly (*P*=.07) or monthly (*P*=.60) assessment attempts. In addition to prior information about patients' medication adherence, patients' age and baseline physical functioning consistently contributed to the predictive power of these models. Assuming weekly assessment attempts, the point on the ROC curve with the greatest number of assessments correctly classified (probability of adherence problems=.58) had a sensitivity of .86 and a specificity of .41.

### Days in Bed

Models predicting days in bed due to mental health problem had the lowest predictive accuracy as measured by the AUC's for models based on weekly, biweekly, and monthly assessment attempts (Table 3 and Figure 4). In Figure 4, the blue line represents the ROC curve for the model based on weekly assessment attempts, the green line represents biweekly assessment attempts, and the red line represents monthly attempts. The yellow line represents the prediction with baseline data only, and all other models also included baseline clinical and sociodemographic information. While the AUC for weekly assessments was significantly different than either biweekly (*P*=.05) or monthly assessments (*P*=.05), there was no statistically significant difference in the AUC for biweekly and monthly calls, (*P*=.57). In addition to the patient's prior reports of days in bed, prior reports of depressive symptoms, as well as their baseline physical and mental functioning were significant predictors of days in bed.

**Figure 1.** Receiver Operator Characteristic (ROC) curves for models predicting patient reports of moderate/severe depression, as measured by a PHQ-9 score ≥10.

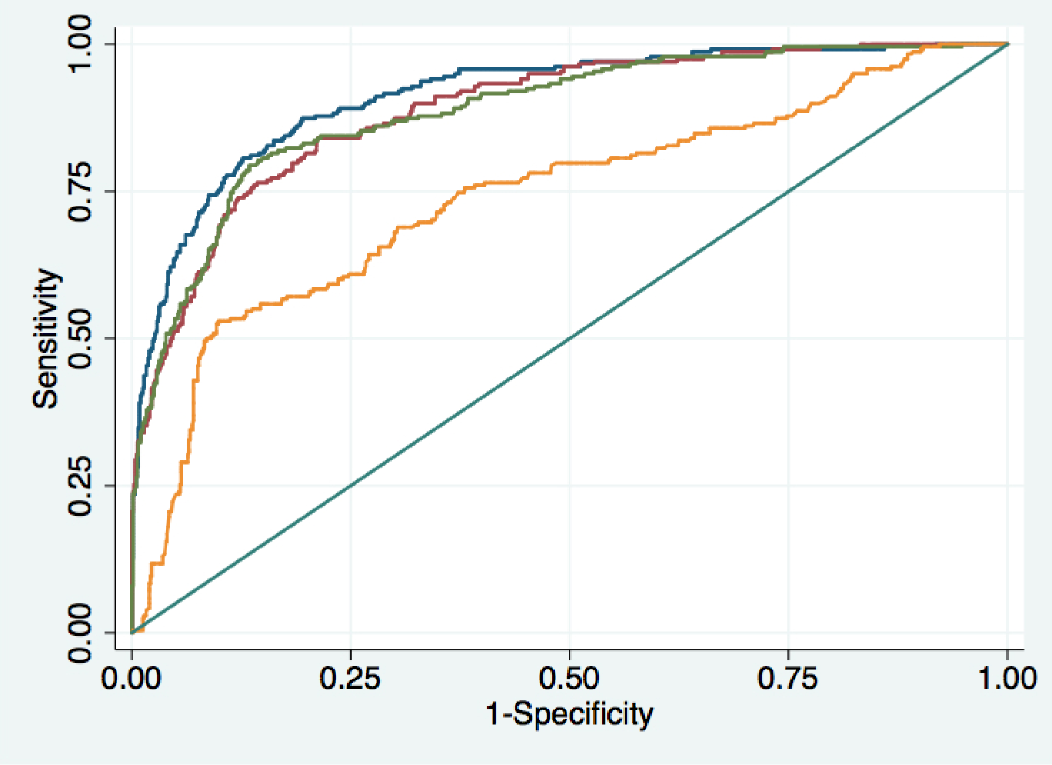


**Figure 2.** Receiver Operator Characteristic (ROC) curves for models predicting patient reports of fair or poor general health status.

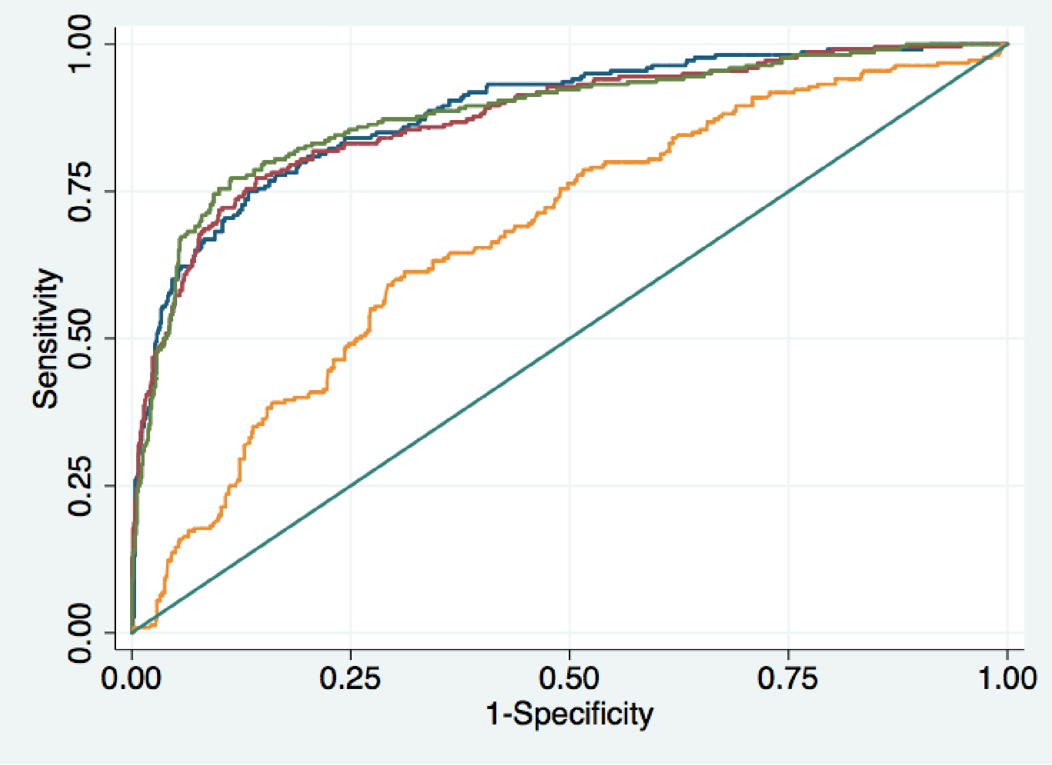


**Figure 3.** Receiver Operator Characteristic (ROC) curves for models predicting patient reports of poor antidepressant medication adherence.

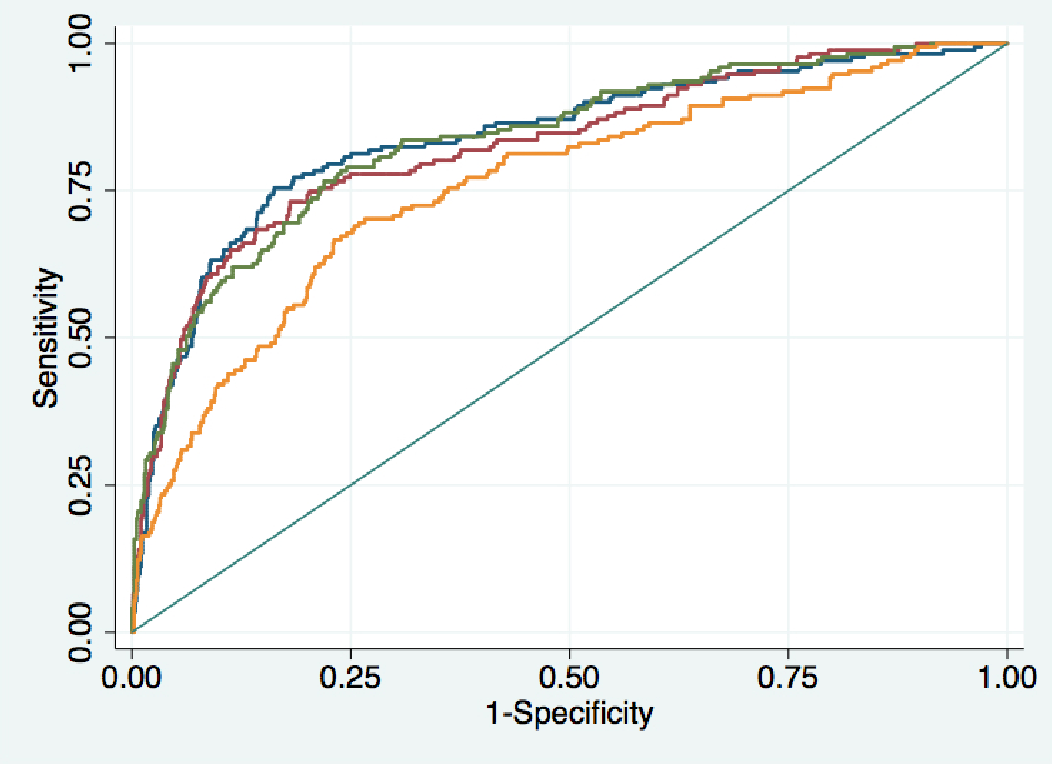
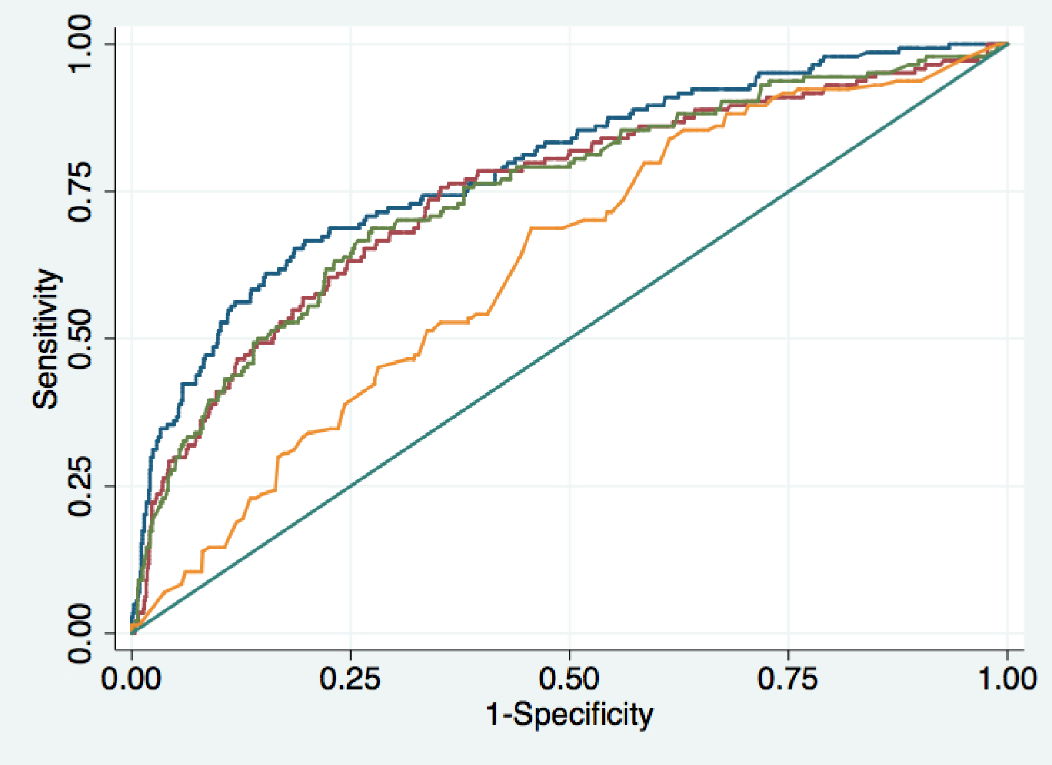


**Figure 4.** Receiver Operator Characteristic (ROC) curves for models predicting patient reports of being bedbound due to mental health problems.



## Discussion

### Principal Findings

These analyses suggest that some IVR assessments of health and behavioral risk factors among patients with depression diagnoses may be unnecessary because patients' responses are predictable based on their prior pattern of reports. In particular, we found that there is little to be gained from asking patients to report their PHQ-9 depression scores weekly and only a negligible incremental difference between biweekly and monthly assessment attempts. A similar pattern was observed with patients' reports of fair or poor perceived general health.

Less frequent assessments of a given health indicator, particularly when that indicator is measured via a multi-item scale such as the PHQ-9, would have two benefits. First, it may be possible to decrease patients' response burden and risk for dropout by avoiding repetitive assessments of the same health problem. Second, by avoiding redundancy in IVR monitoring, more efficient messages could be designed that would cover a broader range of clinical parameters. In the current study, patients reported an average of more than two comorbid chronic conditions. Minimizing redundant questioning would allow for more comprehensive monitoring of comorbidities that may complicate the treatment of patients' depression and pose an independent threat to patients' health.

For two of the outcomes we examined—medication nonadherence and bed-bound status—prior IVR reports were only moderately successful in predicting patients' responses in a subsequent call. Several explanations are possible. It may be that adherence and days in bed were not reliably measured or that other still unmeasured predictors are more important in determining these health behaviors prospectively. Or it may be that these health indicators were in fact changing in unpredictable ways more rapidly than the frequency of monitoring could detect. If the latter reason is true, it may mean that even more frequent assessments are needed to detect all problems that arise. In any case, the approach to examining the frequency of monitoring presented here represents a framework for evaluating those options and making more informed choices about what health indicators to monitor and how often.

Assessments conducted in the current study were completed as part of a clinical service, with feedback to patients' primary care team and informal caregivers when serious problems were reported. It may be that those feedback reports led to interventions that stabilized patients' health status in ways that made subsequent patient reports more predictable. For obvious reasons, collecting patient health information without acting on it would be ethically challenging, but such information could provide insights into the appropriate periodicity of IVR monitoring for various outcomes. On the other hand, data used in the current study are more representative of what patients are likely to report in "real-world" practices, and the fact that we found that weekly assessments may produce redundant information is encouraging for health care organizations struggling with how best to manage their patients with multiple, competing health demands.

Patients who recently changed their antidepressant medication regimen may be more likely to experience side effects leading to adherence problems. The current system was not linked to pharmacy records. Such linkages represent an excellent example of the way in which monitoring systems that include a broader array of potential determinants of patients' health may help ensure that mobile health services focus on health indicators providing the most prognostically important information in the context of everything that is known about the patient.

Predictive models such as these could be used along with advanced machine learning algorithms to tailor the frequency of monitoring across patients, time, and health indicators. For example, time saved gathering redundant information about the trajectory of patients' depressive symptoms could be used to provide cognitive behavioral therapy designed to improve patients' mood by teaching skills such as cognitive restructuring or increased pleasurable activities [32]. Or for patients with depression and comorbid medical disorders, more efficient algorithms could adapt automatically in order to focus on the patient's other diseases, symptoms, or self-care behaviors that need greater attention to promote overall wellness. In brief, data mining approaches illustrated in the current study could be linked with algorithms that automatically update the content of patients' repeated mobile health interactions, maximizing the emphasis on patient education while continually monitoring the health problems that pose the greatest risk to patients' current and future risk for complications.

Each of the four outcomes examined could have been characterized using ordinal or even continuous measures, and the choice of dichotomizing the outcomes may have decreased the models' predictive power. We chose binary outcomes because clinical decisions (eg, whether to call the patient, request a visit, or change a prescription) are often binary, and these logistic models lend themselves to comparison via ROC curves that are familiar to many health care professionals. Nevertheless, data mining includes an increasingly large armamentarium of approaches that could be brought to bear on clinical prediction problems, depending on (for example) the functional form of the outcome, the amount of data available, and whether the relationship of interest is represented by "noisy" data generated from an underlying parametric model.

The current study used logistic regression, cross validation, and ROC curves to identify the predictive trends in patients' IVR-reported data. Artificial neural networks (ANNs) are an alternative parametric approach with more than 15 years of applications to medical diagnostics [33]. Support Vector Machines [34] represent a popular, nonparametric alternative to ANNs [35] for complex classification problems, particularly when the boundaries between groups (eg, between depressed and nondepressed patients) are irregular with respect to predictor variables and sufficient data are available for classification despite noise. Hierarchical latent-variable models (eg, Hidden-Markov Models [36]) could be used to capture underlying latent determinants of depression scores so that medical decisions can be conditioned on that latent information. If a continuous depression score were the outcome, moving average models with exponential smoothing could provide an initial understanding of data trends [37,38]. Other methods for modeling nonstationarities include autoregressive integrating moving averages (ARIMA) models [39] or regression-based forecasting models to extract complex characteristics of time series. More general models for state space representation also could be used to describe the motion of dynamic systems and extract position estimates as well as their derivatives eg, velocities or accelerations) from noisy data sources [40].

Regardless of the analytic approach, it may be that prediction of patients' responses could be improved by including more prior information in the prediction (eg, information from a larger number of prior IVR assessments). In the current study, we attempted to strike a balance between maximizing the predictive accuracy for a given patient, and including in the analyses a large, more representative sample of patients with a sufficient number of assessments (ie, by requiring no more than five prior assessments with at least a 1-week, 2-week, and 1-month gap between each). Similar analyses in the context of data from large health plans may significantly improve the evidence base for clinical decision making.

## Conclusions

In summary, the content and frequency of current mobile health assessments is almost entirely based on a fixed schedule and expert opinion, rather than being individualized based on patients' previously reported status. These analyses indicate that the technical feasibility of gathering high frequency health data may in some instances exceed the clinical benefit of doing so. In particular, weekly or biweekly depressive symptom reports may provide little marginal information regarding how the person is doing relative to collecting that information monthly. Data mining may allow us to detect trends in patient reports that can be used by intelligent systems to accurately predict patients' health status. The next generation of automated health assessment services should use these or other data mining techniques to avoid redundant assessments and gather data at the frequency that maximizes the value of the information collected. Such adaptive systems could be much more patient-friendly and could accommodate a much broader set of risk factors for the large and growing number of patients who have multiple chronic diseases.

## Acknowledgments

John Piette is a VA Senior Research Career Scientist. The current study was supported by the Department of Veterans Affairs Health Services Research and Development Program. Other financial support came from the University of Michigan Health System Faculty Group Practice, grant number P30DK092926 from the National Institute of Diabetes and Digestive and Kidney Diseases, grant number IIS 1064948 from the National Science Foundation, and grants number HFP 83-014 and 11-088 from the VA Health Services Research and Development Program. Diana Parrish, Dana Striplin, and Nicolle Marinec played essential roles in the design and implementation of this mobile health depression management program. Steven Bernstein provided important insights into the implications of repeated health monitoring in clinical practice.

## Conflicts of Interest

None declared.

## References

1. Kahn JG, Yang JS, Kahn JS. 'Mobile' health needs and opportunities in developing countries. Health Aff (Millwood) 2010 Feb;29(2):252-258 [FREE Full text] [doi: 10.1377/hlthaff.2009.0965] [Medline: 20348069]
2. Milne RG, Horne M, Torsney B. SMS reminders in the UK national health service: an evaluation of its impact on "no-shows" at hospital out-patient clinics. Health Care Manage Rev 2006;31(2):130-136. [Medline: 16648692]
3. Barclay E. Text messages could hasten tuberculosis drug compliance. Lancet 2009 Jan 3;373(9657):15-16. [Medline: 19125443]
4. Moore HK, Mundt JC, Modell JG, Rodrigues HE, DeBrota DJ, Jefferson JJ, et al. An examination of 26,168 Hamilton Depression Rating Scale scores administered via interactive voice response across 17 randomized clinical trials. J Clin Psychopharmacol 2006 Jun;26(3):321-324. [doi: 10.1097/01.jcp.0000219918.96434.4d] [Medline: 16702899]
5. González GM, Costello CR, La Tourette TR, Joyce LK, Valenzuela M. Bilingual telephone-assisted computerized speech-recognition assessment: is a voice-activated computer program a culturally and linguistically appropriate tool for screening depression in English and Spanish? Cult Divers Ment Health 1997;3(2):93-111. [Medline: 9231537]
6. Kobak KA, Taylor LH, Dottl SL, Greist JH, Jefferson JW, Burroughs D, et al. A computer-administered telephone interview to identify mental disorders. JAMA 1997 Sep 17;278(11):905-910. [Medline: 9302242]
7. Mundt JC, Kobak KA, Taylor LV, Mantle JM, Jefferson JW, Katzelnick DJ, et al. Administration of the Hamilton Depression Rating Scale using interactive voice response technology. MD Comput 1998;15(1):31-39. [Medline: 9458661]
8. Brodey BB, Rosen CS, Winters KC, Brodey IS, Sheetz BM, Steinfeld RR, et al. Conversion and validation of the Teen-Addiction Severity Index (T-ASI) for Internet and automated-telephone self-report administration. Psychol Addict Behav 2005 Mar;19(1):54-61. [doi: 10.1037/0893-164X.19.1.54] [Medline: 15783278]
9. Mundt JC, Moore HK, Bean P. An interactive voice response program to reduce drinking relapse: a feasibility study. J Subst Abuse Treat 2006 Jan;30(1):21-29. [doi: 10.1016/j.jsat.2005.08.010] [Medline: 16377449]
10. Bopp JM, Miklowitz DJ, Goodwin GM, Stevens W, Rendell JM, Geddes JR. The longitudinal course of bipolar disorder as revealed through weekly text messaging: a feasibility study. Bipolar Disord 2010 May;12(3):327-334 [FREE Full text] [doi: 10.1111/j.1399-5618.2010.00807.x] [Medline: 20565440]
11. Bauer S, Moessner M. Technology-enhanced monitoring in psychotherapy and e-mental health. J Ment Health 2012 Aug;21(4):355-363. [doi: 10.3109/09638237.2012.667886] [Medline: 22548363]
12. Piette JD, Marinec N, Gallegos-Cabriales EC, Gutierrez-Valverde JM, Rodriguez-Saldaña J, Mendoz-Alevares M, et al. Spanish-speaking patients' engagement in interactive voice response (IVR) support calls for chronic disease self-management: data from three countries. J Telemed Telecare 2013 Mar 26 (forthcoming). [doi: 10.1177/1357633X13476234] [Medline: 23532005]

13. Piette JD, Rosland AM, Marinec NS, Striplin D, Bernstein SJ, Silveira MJ. Engagement with automated patient monitoring and self-management support calls: experience with a thousand chronically ill patients. Med Care 2013 Mar;51(3):216-223. [doi: 10.1097/MLR.0b013e318277ebf8] [Medline: 23222527]
14. Piette JD. Patient education via automated calls: a study of English and Spanish speakers with diabetes. Am J Prev Med 1999 Aug;17(2):138-141. [Medline: 10490057]
15. Piette JD, McPhee SJ, Weinberger M, Mah CA, Kraemer FB. Use of automated telephone disease management calls in an ethnically diverse sample of low-income patients with diabetes. Diabetes Care 1999 Aug;22(8):1302-1309 [FREE Full text] [Medline: 10480775]
16. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. Lancet 2012 Jul 7;380(9836):37-43. [doi: 10.1016/S0140-6736(12)60240-2] [Medline: 22579043]
17. Fortin M, Hudon C, Haggerty J, Akker M, Almirall J. Prevalence estimates of multimorbidity: a comparative study of two sources. BMC Health Serv Res 2010;10:111 [FREE Full text] [doi: 10.1186/1472-6963-10-111] [Medline: 20459621]
18. Lichtenstein MJ, Sweetnam PM, Elwood PC. Visit frequency for controlled essential hypertension: general practitioners' opinions. J Fam Pract 1986 Oct;23(4):331-336. [Medline: 3760794]
19. Schectman G, Barnas G, Laud P, Cantwell L, Horton M, Zarling EJ. Prolonging the return visit interval in primary care. Am J Med 2005 Apr;118(4):393-399. [doi: 10.1016/j.amjmed.2005.01.003] [Medline: 15808137]
20. DeSalvo KB, Block JP, Muntner P, Merrill W. Predictors of variation in office visit interval assignment. Int J Qual Health Care 2003 Oct;15(5):399-405 [FREE Full text] [Medline: 14527983]
21. DeSalvo KB, Bowdish BE, Alper AS, Grossman DM, Merrill WW. Physician practice variation in assignment of return interval. Arch Intern Med 2000 Jan 24;160(2):205-208. [Medline: 10647759]
22. Cios KJ, Kacprzyk J. Medical Data Mining and Knowledge Discovery. New York, NY: Physica-Verlag; 2001.
23. Cios KJ, Moore GW. Uniqueness of medical data mining. Artif Intell Med 2002;26(1-2):1-24. [Medline: 12234714]
24. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [Medline: 11556941]
25. Löwe B, Unützer J, Callahan CM, Perkins AJ, Kroenke K. Monitoring depression treatment outcomes with the patient health questionnaire-9. Med Care 2004 Dec;42(12):1194-1201. [Medline: 15550799]
26. Miilunpalo S, Vuori I, Oja P, Pasanen M, Urponen H. Self-rated health status as a health measure: the predictive value of self-reported health status on the use of physician services and on mortality in the working-age population. J Clin Epidemiol 1997 May;50(5):517-528. [Medline: 9180644]
27. Grant MD, Piotrowski ZH, Chappell R. Self-reported health and survival in the Longitudinal Study of Aging, 1984-1986. J Clin Epidemiol 1995 Mar;48(3):375-387. [Medline: 7897459]
28. DeSalvo KB, Bloser N, Reynolds K, He J, Muntner P. Mortality prediction with a single general self-rated health question. A meta-analysis. J Gen Intern Med 2006 Mar;21(3):267-275 [FREE Full text] [doi: 10.1111/j.1525-1497.2005.00291.x] [Medline: 16336622]
29. Kroenke K, Strine TW, Spitzer RL, Williams JB, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009 Apr;114(1-3):163-173. [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]
30. Ware J, Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med Care 1996 Mar;34(3):220-233. [Medline: 8628042]
31. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 1997 Jul;30(7):1145-1159. [doi: 10.1016/S0031-3203(96)00142-2]
32. Naylor MR, Helzer JE, Naud S, Keefe FJ. Automated telephone as an adjunct for the treatment of chronic pain: a pilot study. J Pain 2002 Dec;3(6):429-438. [Medline: 14622728]
33. Armoni A. Use of neural networks in medical diagnosis. MD Comput 1998;15(2):100-104. [Medline: 9540322]
34. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995 Sep;20(3):273-297. [doi: 10.1007/BF00994018]
35. Tonello L, Vescini F, Caudarella R. Support vector machines versus artificial neural network: Who is the winner? Kidney Int 2007 Jan;71(1):84-85. [doi: 10.1038/sj.ki.5001907] [Medline: 17167512]
36. Fine S, Singer Y, Tishby N. The hierarchical hidden marcov model: analysis and applications. Machine Learning 1998;32:41-62.
37. Spliid H. Monitoring medical procedures by exponential smoothing. Stat Med 2007 Jan 15;26(1):124-138. [doi: 10.1002/sim.2520] [Medline: 16479560]
38. Verrier RL. Elevated T-wave alternans predicts nonsustained ventricular tachycardia in association with percutaneous coronary intervention in ST-segment elevation myocardial infarction (STEMI) patients. Journal of Cardiovascular Electophysiology 2013 (forthcoming).
39. Sebestyen B, Rihmer Z, Balint L, Szokontor N, Gonda X, Gyarmati B, et al. Gender differences in antidepressant use-related seasonality change in suicide mortality in Hungary, 1998-2006. World J Biol Psychiatry 2010 Apr;11(3):579-585. [doi: 10.3109/15622970903397722] [Medline: 20218927]
40. Lavieri MS, Puterman L, Tyldesley S, Morris WJ. When to treat prostate cancer patients based on their PSA dynamics. IIE Transactions on Healthcare Systems Engineering 2012;2:62-77.

## Abbreviations

**ANN:** artificial neural networks
**ARIMA:** autoregressive integrating moving averages
**AUC:** area under the curve
**IVR:** interactive voice response calls
**PHQ-9:** 9-item Patient Health Questionnaire
**ROC:** Receiver Operator Characteristic curve

*Edited by G Eysenbach; submitted 01.03.13; peer-reviewed by H Baloch, M Meglic; comments to author 30.03.13; revised version received 23.04.13; accepted 23.04.13; published 05.07.13*

*Please cite as:*
*Piette JD, Sussman JB, Pfeiffer PN, Silveira MJ, Singh S, Lavieri MS*
*Maximizing the Value of Mobile Health Monitoring by Avoiding Redundant Patient Reports: Prediction of Depression-Related Symptoms and Adherence Problems in Automated Health Assessment Services*
*J Med Internet Res 2013;15(7):e118*
*URL: http://www.jmir.org/2013/7/e118/*
*doi: 10.2196/jmir.2582*
*PMID: 23832021*