

# Voice Response System to Measure Healthcare Costs: A STAR\*D Report

T. Michael Kashner, PhD, JD; Madhukar H. Trivedi, MD; Annie Wicker, BS; Maurizio Fava, MD;  
John H. Greist, MD; James C. Mundt, PhD; Kathy Shores-Wilson, PhD;  
A. John Rush, MD; and Stephen R. Wisniewski, PhD

Provider records often are considered the gold standard to monitor patient service histories, measure use of care, and to compute healthcare costs. However, accessing such records from diverse healthcare providers and in the wake of Health Insurance Portability and Accountability Act of 1996 (HIPAA) regulations can present many challenges, as files often are inaccessible<sup>1</sup> and vary by format, completeness, and data accuracy.<sup>2</sup> Thus, plan administrators, policy makers, and research investigators often turn to self-reported data when provider records are not available.<sup>3</sup>

In this study, we evaluate the performance of an interactive voice response (IVR) system that collected healthcare utilization and costs information from a computerized script administered by phone (UAC-IVR) for the Sequenced Treatment Alternatives to Relieve Depression (STAR\*D) study.<sup>4-8</sup> STAR\*D followed approximately 4000 patients who were being treated for nonpsychotic major depressive disorder (MDD) by 400 clinicians at 41 sites in both specialty and primary care settings, and in both the public and private sectors. To capture use-of-care information, study subjects were asked to dial a centralized number, listen to instructions, and answer computer-scripted questions by pressing keys on a touch-tone telephone pad.

As a data collection tool, IVR systems have gained both public<sup>9-12</sup> and clinical<sup>13,14</sup> acceptance. Compared with personal interviews, IVR responses are associated with lower collection costs, greater patient convenience, and fewer transcription errors.<sup>15</sup> These systems also allow for remote data access, automated scoring,<sup>16</sup> patient feedback,<sup>17</sup> and opportunities for self-disclosure of sensitive information.<sup>16,18-21</sup> Furthermore, IVR technology has been applied to studies on alcohol use,<sup>22,23</sup> cognitive functioning,<sup>24</sup> work and social adjustment,<sup>25</sup> chronic insomnia,<sup>26</sup> smoking cessation,<sup>27</sup> depressive symptoms,<sup>28</sup> and obsessive-compulsive disorder.<sup>29</sup> Good reliability has been reported when IVR results are compared with responses from written questionnaires and personal interviews.<sup>30</sup> A high correspondence has been found between psychiatric diagnoses based on the Primary Care Evaluation of Mental Disorders (PRIME-MD) screening instrument using IVR technology and those obtained using the Structured Clinical Interview for DSM-IV (SCID-IV) interview.<sup>31</sup> The specificity and sensitivity of an IVR mental health screener for identifying anxiety and depressive disorders, obsessive-compulsive dis-

**Objective:** To evaluate a telephone-operated, interactive voice response (IVR) system designed to collect use-of-care data from patients with major depression (UAC-IVR).

**Study Design:** Patient self-reports from repeated IVR surveys were compared with provider records for 3789 patients with major depression at 41 clinical sites participating in the Sequenced Treatment Alternatives to Relieve Depression (STAR\*D) trial.

**Methods:** UAC-IVR responses were examined for consistency and compared with provider records to compute reporting biases and intraclass correlation coefficients. Predictors of inconsistent responses and reporting biases were based on mixed logistic and regression models adjusted for need and predisposing and enabling covariates, and corrected for nesting and repeated measures.

**Results:** Inconsistent responses were found for 10% of calls and 21% of patients. Under-reporting biases (–20%) and moderate agreement (intraclass correlation of 68%) were found when UAC-IVR responses were compared with medical records. IVR reporting biases were less for patients after 3 calls or more (experience), for patients with severe baseline symptoms (motivation), and for patients who gave consistent IVR responses (reliability). Bias was unrelated to treatment outcomes or demographic factors.

**Conclusion:** Clinical managers should use IVR systems to collect service histories only after patients are properly trained and responses monitored for consistency and reporting biases.

(*Am J Manag Care.* 2009;15(3):153-162)

**In this issue**  
Take-Away Points / p160  
[www.ajmc.com](http://www.ajmc.com)  
Full text and PDF

**For author information and disclosures,  
see end of text.**

orders, eating disorders, and alcohol use disorders also have been demonstrated.<sup>32</sup>

Prior studies have not focused on IVRs as a data collection tool to measure patient total use of care. In this study, we evaluated STAR\*D's new use-of-care survey, UAC-IVR, for both consistency and reliability. Consistency was determined by comparing responses to questions that asked if any care was used (yes/no) with questions that asked how much care was used (greater than zero/none). Reliability was assessed by comparing survey responses with provider records.

## METHODS

### Data

The STAR\*D consent protocol<sup>1</sup> and study design<sup>4,8</sup> are described elsewhere. Briefly, subjects signed an institutional review board–approved informed consent form and were followed through prospective and sequenced treatments for MDD. Patients who responded to treatment or achieved remission were followed for an additional year. Data were collected from both providers and patients. Patient information was solicited from written questionnaires, face-to-face and telephone interviews, and patient-initiated IVR calls administered by Healthcare Technology Systems, Inc. in Madison, Wisconsin. STAR\*D research staff helped patients make calls at baseline, after 6 weeks at each treatment level, at the end of each treatment level, at monthly intervals during the 12-month follow-up, and at study exit. To make a call, patients first dialed a toll-free number using a touch-tone telephone. The caller received recorded instructions, followed by a set of questions. After each question, the recorded message prompted patients to respond by pressing an appropriate number on the telephone keypad. The computer then recorded each response and automatically determined the next set of scripted questions to ask the patient.

Scripted questions covering patient use of healthcare during 90-day intervals are presented in the **Figure**. Questions were derived from the Utilization and Cost Methodology (UAC).<sup>3,33-36</sup> Each time subjects accessed the IVR server, the computer checked to see whether the use-of-care script had ran within the past 90 days. This strategy minimized risks of double-counting services from overlapping observation periods between IVR calls. Periods not covered by an IVR call were treated as missing.

Respondents were first asked whether they had used care during the past 3 months (yes or no). Patients who responded “yes” were subsequently asked how much care they had used. Patients were asked about using services classified by setting (outpatient clinic visits, emergency room visits, and inpatient days stayed) and by type (depression related, other-psychiatric, and general medical problems). For evaluative purposes, a

response was considered “inconsistent” if the respondent answered “yes” to using care while subsequently reporting that “zero” days or visits were actually used. Responses that were not inconsistent were considered consistent.

### Other Data

To assess reliability, provider data were obtained from billing claims and medical charts for patients signing a medical release. STAR\*D focused on depression-related care; thus, records for services not related to depression were generally unavailable, and these analyses were limited to depression-related outpatient visits only. Services were classified by *Current Procedural Terminology* (CPT),<sup>37</sup> the level 1 Healthcare Common Procedure Coding System,<sup>38</sup> and psychiatric diagnoses based on the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV)*.<sup>39</sup> Abstractors counted the number of outpatient visits that were depression related (DSM-IV 296, 311) but not emergency room related (CPT 99281, 99282, 99283, 99284, 99285, 99288) for each 90-day period ending on the date of the respective IVR survey.

Patient demographic, education, earnings, employment, and health insurance information were taken from face-to-face and telephone interviews. Patients rated (agreed or neutral/disagreed): “If I can get the help I need from a doctor, I believe that I will be much better able: (a) to make important decisions that affect my life and those of my family? and (b) to enjoy things that interest me?” Patients also rated (helpful vs neutral/not helpful): “the current overall impact of your family and friends on your condition?” Also collected using the IVR system were family size, health insurance status, Medicaid eligibility, and mental and physical functioning based on the Medical Outcome Study 12-item short form.<sup>40</sup>

For purposes of these analyses, treatment outcomes were based on the 17-item Hamilton Rating Scale for Depression (HRSD<sub>17</sub>)<sup>41,42</sup> administered by telephone at the end of the first treatment step (citalopram), and on the written, patient self-reported, 16-item Quick Inventory of Depressive Symptomatology (QIDS-SR<sub>16</sub>),<sup>43-46</sup> administered at baseline and at each STAR\*D clinic visit. Treatment outcomes were computed as (1) remission defined by an HRSD<sub>17</sub> score of 7 or lower at exit from the first treatment step, (2) remission defined by a QIDS-SR<sub>16</sub> score of 5 or lower, and (3) response defined by a reduction in QIDS-SR<sub>16</sub> score from baseline of 50% or more. Patients with missing HRSD<sub>17</sub> scores were not considered to have achieved remission.<sup>4</sup>

## ANALYSES

Interactive voice response use-of-care responses were evaluated for both consistency and reliability. To assess reliabil-



ity, counts of depression-related, nonemergency outpatient visits based on IVR responses were compared with provider records that spanned comparable time periods to compute bias (mean difference) and intraclass correlation from 2-way mixed models.<sup>47,48</sup>

The associations of selected predictor variables with response inconsistency were computed from 3-level mixed logistic models. Similarly, the associations of selected predictor variables with reporting biases were computed from 3-level mixed regression models. Both sets of models were computed using Hierarchical Linear Modeling software,<sup>49</sup> where level 1 is IVR calls, level 2 is individual patients, and level 3 is study sites. Both models corrected for facility nesting and repeated measures with random-effects terms. Estimates of the association for each selected predictor variable were adjusted, in turn, for the mean-centered values of a given set of covariates. Based on traditional theory,<sup>50</sup> these covariates included need (baseline QIDS-SR<sub>16</sub>, age), predisposing (graduated high school, Hispanic, African American, sex), and enabling (married, enrolled in private health insurance plan, and employed) variables. An additional covariate was added: the order of the call. To account for complex error distributions, significance tests were based on robust estimates of standard errors.

## RESULTS

### Sample Features

STAR\*D enrolled 4041 subjects, of whom 94% (n = 3789) completed an IVR use-of-care script, making 9864 calls, or 2.6 calls per patient (SD = 1.7, range = 1-8) (Table 1). Among

■ **Table 1.** Number of IVR Calls per Patient When the Use-of-Care Script Ran<sup>a</sup>

Number of Calls	Number (%) of Patients
1 or more	3789 (100.0)
1	1468 (38.7)
2	738 (19.5)
3	465 (12.3)
4	358 (9.4)
5	501 (13.2)
6	226 (6.0)
7	32 (0.8)
8	1 (0.0)

IVR indicates interactive voice response.

<sup>a</sup>A completed use-of-care IVR call was a scheduled call that had been initiated by the respondent, the scripted use-of-care questions were asked, and the respondent provided answers to all use-of-care questions.

the 3789 subjects who completed the script, 17% (655 of 3789) were African American (excluding Hispanic black); 12% (464 of 3784) were Hispanic; 63% (2377 of 3788) were female; 34% (1270 of 3785) were married; 57% (2157 of 3784) were employed; 88% (3315 of 3784) had a high school diploma, General Educational Development (GED) certification, or higher; 12% (458 of 3691) had Medicaid coverage; and 51% (1911 of 3715) had private health insurance. The mean age was 41.2 years (SD = 13.2 years).

### Inconsistent Responses

Use-of-care information obtained from IVR calls is summarized in Table 2. All patients reported using depression-related care on at least 1 call. For all calls activating the use-of-care script, 50% (4921 of 9864) reported depression-related care, 36% (3545 of 9864) reported general medical care, 12% (1137 of 9864) reported other psychiatric care, 2% (166 of 9864) reported an inpatient stay for depression, and 3% (311 of 9864) reported an inpatient stay for general medical purposes.

Table 2 also lists by setting the number of sites, subjects, and IVR calls that contained inconsistent use-of-care responses. There were 1069 instances of an inconsistent response, among 944 of 9864 (10%) IVR calls, from 778 of 3789 (21%) participants, at 39 of 41 (95%) study sites. By comparison, 14% (537 of 3745) of patients gave inconsistent employment and earnings responses.

Among the 944 calls with at least 1 inconsistent use-of-care response, 832 (88%) involved only 1 response, 99 (10%) involved 2 responses, and 13 (1%) involved 3 inconsistent responses during the same call. Among the 778 participants who gave inconsistent responses, 648 (83%) made 1 call, 99 (13%) made 2 calls, 26 (3%) made 3 calls, and 5 (1%) made 4 or more calls involving at least 1 inconsistent response.

Adjusted odds ratios for predictors of inconsistent response by IVR call are listed in Table 3. The likelihood of making an inconsistent response tended to decline as patients gained experience with IVR calling. Whether this trend reflects patient learning or sample selection biases cannot be determined. Patients who consented to release their medical records to study investigators were slightly less likely to provide inconsistent responses, though the effect was not statistically significant. Patients providing inconsistent employment and earnings data were neither more nor less likely to provide inconsistent use-of-care responses on their IVR calls.

There were no significant associations between treatment outcomes (response or remission) and the likelihood of giving inconsistent use-of-care responses. However, patients were actually less likely to give inconsistent responses when they reported more severe baseline depressive symptoms or poorer

■ **Table 2.** Number of Completed IVR Calls by Setting, Disorder, and Consistency Status<sup>a</sup>

Call Characteristics	All Settings	Outpatient Care	Emergency Room	Inpatient Care
<b>Total</b>				
Subjects	3789	3153	1154	538
Calls	9864	6771	1453	634
<b>By disorder</b>				
<b>Depression-related calls<sup>b</sup></b>	4921 (50%)	4729 (48%)	446 (5%)	166 (2%)
Mean ± SD volume per call <sup>c</sup>		4.1 ± 4.8	1.4 ± 1.9	5.2 ± 5.0
<b>Other-psychiatric calls<sup>b</sup></b>	1137 (12%)	1048 (11%)	132 (1%)	42 (0%)
Mean ± SD volume per call <sup>c</sup>		3.2 ± 5.3	1.5 ± 0.9	6.4 ± 7.5
<b>General-medical calls<sup>b</sup></b>	3545 (36%)	3280 (33%)	937 (9%)	311 (3%)
Mean ± SD volume per call <sup>c</sup>		3.0 ± 3.8	1.7 ± 1.5	5.6 ± 9.1
<b>With inconsistent response</b>				
Sites <sup>d</sup>	39 (95%)	39 (95%)	35 (85%)	33 (89%)
Subjects <sup>d</sup>	778 (21%)	575 (18%)	221 (19%)	149 (28%)
Calls <sup>d</sup>	944 (10%)	684 (10%)	226 (16%)	159 (25%)

IVR indicates interactive voice response; STAR\*D, Sequenced Treatment Alternatives to Relieve Depression.

<sup>a</sup>IVR completed calls were initiated by respondents who completed the utilization of care questions.

<sup>b</sup>Number of completed calls reporting service use (percentage of total completed calls).

<sup>c</sup>Mean volume per IVR call expressed in terms of visits (outpatient care, emergency room) or days (inpatient care) when nonzero use was indicated.

<sup>d</sup>Number of sites, number of subjects, and number of completed IVR calls with at least 1 inconsistent use-of-care response (percentage of total by treatment setting). There were 41 STAR\*D sites. An inconsistent response occurred when the patient reported zero volume after indicating some care had been received, by service category.

mental functioning. On the other hand, patients with worse physical functioning were more likely to provide inconsistent responses. The different roles for mental versus physical health on response consistencies suggest that psychiatric symptoms may be motivating respondents to complete study surveys accurately, whereas poor physical functioning may be related to the capacity of patients to respond.

Patient income, employment, insurance status, and health attitudes were not related to response consistency. On the other hand, respondents who were older, female, Hispanic, African American, less educated, or married were more likely to make inconsistent choices on any given IVR call than their counterparts.

### Reliability

Interactive voice response reporting biases for outpatient depression-related care were determined by subtracting counts computed with IVR responses from counts based on medical records. Estimates were taken from 6858 complete and consistent IVR calls representing 2677 participants for whom medical records were available at 39 clinic sites. Overall, IVR responses underreported depression-related treatment visits by 19.2%: 2.1 visits with IVR versus 2.6 visits with the medical record ( $\Delta = -0.50$  visits  $\pm .045$ ; 95% confidence interval [CI],  $-0.59, -0.42$ ;  $t = 11.14$ ;  $df = 6857$ ;

$P < .001$ ) with an intraclass correlation of 0.49, based on a 2-way mixed-effects model. Adjusting for core factors and facility nesting, IVR underreporting bias was 34.2% (adjusted  $\Delta = -0.89$  visits  $\pm 0.12$ ; 95% CI,  $-0.66, -1.13$ ;  $t = 7.52$ ;  $df = 2566$ ;  $P < .001$ ). Computing total visits across all IVR calls, respondents understated visits compared with provider records by 19.2%: 5.31 visits with IVR versus 6.60 visits with the medical record ( $\Delta = -1.29$  visits  $\pm 0.14$ ; 95% CI,  $-1.57, -1.02$ ;  $t = 9.27$ ;  $df = 2676$ ;  $P < .001$ ), with an intraclass correlation of 0.68.

Selected predictors of IVR reporting biases are listed in **Table 4**. A positive effect means the predictor was associated with less bias because counts of outpatient visits computed from IVR responses understated counts computed from medical records. Conversely, negative effects reflect greater disparities in visit counts between survey responses and medical records.

Interactive voice response biases tended to get worse for patients who used more care (based on medical records), made inconsistent responses on other IVR calls, had less severe baseline symptoms for depression, were uninsured, and were not on sick leave. On the other hand, there was no evidence that reporting biases were associated with treatment outcomes or patient demographic characteristics, income, employment status, or attitudes about care. Overall, these data suggest patients may be more careful when responding

**Table 3.** Adjusted<sup>a</sup> Association Between Selected Predictors of Inconsistent Responses<sup>b</sup> to Use-of-Care Questions for Completed IVR Calls<sup>c</sup>

Predictor	Odds Ratio (95% CI)	Wald <i>t</i>	<i>P</i>
<b>Administrative status</b>			
Order of IVR call	0.64 (0.56, 0.74)	6.37	<.001
Order of IVR call—squared	1.06 (1.04, 1.09)	4.89	<.001
Medical record release signed	0.86 (0.71, 1.05)	1.44	.15
Inconsistent earnings response	1.01 (0.80, 1.26)	0.05	.96
<b>Level 1 outcome</b>			
QIDS-SR <sub>16</sub> response <sup>d</sup>	1.13 (0.95, 1.35)	1.36	.17
QIDS-SR <sub>16</sub> remission <sup>e</sup>	1.05 (0.87, 1.27)	0.49	.62
HRSD <sub>17</sub> remission <sup>f</sup>	1.06 (0.88, 1.27)	0.60	.55
<b>Health status</b>			
Baseline QIDS-SR <sub>16</sub>	0.98 (0.97, 0.99)	2.92	.004
Family history of depression	1.02 (0.90, 1.16)	0.38	.70
Mental functioning: 10-point difference	1.23 (1.13, 1.34)	4.77	<.001
Physical functioning: 10-point difference	0.88 (0.83, 0.94)	4.06	<.001
On sick leave from work	0.94 (0.75, 1.17)	0.84	.40
<b>Income/earnings</b>			
Household income: \$10k difference	0.95 (0.63, 1.42)	0.25	.80
Earned income: \$10k difference	0.93 (0.60, 1.44)	0.32	.75
Employment status	0.89 (0.77, 1.02)	1.66	.097
Volunteer status	1.02 (0.85, 1.22)	0.19	.85
<b>Insurance status</b>			
Medicaid	1.13 (0.94, 1.36)	1.27	.21
Private insurance	0.99 (0.82, 1.20)	0.09	.95
<b>Perception/attitudes</b>			
Care helpful in decision-making <sup>g</sup>	0.92 (0.72, 1.17)	0.66	.51
Care helpful to enjoy things <sup>g</sup>	1.02 (0.79, 1.33)	0.18	.86
Family helpful <sup>h</sup>	1.07 (0.93, 1.23)	0.90	.37
<b>Demographic characteristics</b>			
Age: 10-y difference	1.20 (1.13, 1.26)	6.59	<.001
Male	0.79 (0.69, 0.90)	3.49	.001
Hispanic	1.24 (1.01, 1.53)	2.08	.037
African American	1.85 (1.59, 2.16)	7.94	<.001
High school graduate or higher	0.54 (0.40, 0.71)	4.28	<.001
Student status	0.92 (0.73, 1.16)	0.71	.48
Married	1.15 (1.03, 1.28)	2.48	.013
Living with spouse	1.13 (0.86, 1.50)	0.87	.39
Household size	1.01 (0.98, 1.03)	0.60	.55
Lived at residence >10-y	1.15 (0.92, 1.44)	1.23	.22

CI indicates confidence interval; HRSD<sub>17</sub>, 17-item Hamilton Rating Scale for Depression; IVR, interactive voice response; QIDS-SR<sub>16</sub>, 16-item Quick Inventory of Depressive Symptomatology.

<sup>a</sup>Estimates were adjusted for covariates: patient age, Hispanic and African American status (patients could indicate both), sex, marital status (married vs not married for all reasons), education (high school graduate or higher vs less than high school graduate), private insurance status (yes/no), paid employment (full or part time vs unemployed for all reasons), baseline QIDS-SR<sub>16</sub> score, and order of IVR call where 1 indicates the first completed call, 2 indicates the second completed call, and so forth. If the selected predictor also was a covariate, the term was entered only once in the final model.

<sup>b</sup>An inconsistent IVR response to a particular service category occurred when the respondent reported zero volume to that service category after answering “yes” to a prior question indicating some care in that service category had been received.

<sup>c</sup>For 9864 IVR calls initiated by 3789 patients who completed the utilization-of-care questions.

<sup>d</sup>Symptom response defined by at least a 50% reduction in baseline scores on the QIDS-SR<sub>16</sub>—self-report at exit.

<sup>e</sup>Symptom remission defined by a score of 5 or lower on the QIDS-SR<sub>16</sub>—self-report at exit.

<sup>f</sup>Symptom remission defined by a score of 7 or lower on the HRSD<sub>17</sub> at exit.

<sup>g</sup>Patients strongly agree or agree (rather than neutral, disagree, or strongly disagree) that if they get the help they need from a doctor, they will be better able to make important decisions affecting life, or they will be better able to enjoy things.

<sup>h</sup>Patients find the impact of family and friends on their health condition to be very helpful, moderately helpful, or minimally helpful (rather than neutral, minimally, moderately more, or much more difficult).

**Table 4.** Adjusted<sup>a</sup> Predictors of Reporting Biases<sup>b</sup> Comparing 90-Day Depression-Related Outpatient Visits Reported on an IVR Call<sup>c</sup> With Provider Records

Predictor	Bias in Visits	95% Confidence Interval	t Statistic	P
<b>Administrative status</b>				
Order of IVR call	-1.04	-1.36, -0.72	6.44	<.001
Order of IVR call—squared	0.21	0.16, 0.25	9.27	<.001
Recorded number of visits	-0.51	-0.65, -0.36	6.89	<.001
Recorded number of visits—squared	0.00	-0.01, 0.01	-0.14	.89
Inconsistent use response	-0.20	-0.37, -0.03	2.28	.023
Inconsistent earnings response	-0.14	-0.41, 0.14	0.97	.34
<b>Level 1 outcome</b>				
QIDS-SR <sub>16</sub> response <sup>d</sup>	-0.08	-0.32, 0.15	0.67	.50
QIDS-SR <sub>16</sub> remission <sup>e</sup>	-0.06	-0.26, 0.14	0.62	.54
HRSD <sub>17</sub> remission <sup>f</sup>	-0.07	-0.25, 0.11	0.78	.44
<b>Health status</b>				
Baseline QIDS-SR <sub>16</sub>	0.04	0.02, 0.06	4.28	<.001
Family history of depression	0.05	-0.13, 0.22	0.53	.599
Mental functioning: 10-point difference	-0.06	-0.19, 0.07	0.91	.365
Physical functioning: 10-point difference	-0.09	-0.21, 0.03	1.41	.158
On sick leave from work	0.52	0.22, 0.82	3.37	.001
<b>Income/earnings</b>				
Household income: \$10k difference	0.27	-0.18, 0.73	1.17	.24
Earned income: \$10k difference	0.06	-0.29, 0.42	0.35	.73
Employment status	-0.20	-0.43, 0.02	1.79	.073
Volunteer status	0.00	-0.24, 0.23	.03	.98
<b>Insurance status</b>				
Medicaid	-0.07	-0.30, 0.16	0.63	.53
Private insurance	0.36	0.17, 0.56	3.59	.001
<b>Perception/attitudes</b>				
Care helpful in decision-making <sup>g</sup>	-0.15	-0.50, 0.21	0.812	.42
Care helpful to enjoy things <sup>g</sup>	-0.19	-0.61, 0.23	0.878	.38
Family helpful <sup>h</sup>	0.04	-0.15, 0.23	0.407	.68
<b>Demographic characteristics</b>				
Age: 10-y difference	0.01	-0.08, 0.10	0.24	.81
Male	0.18	-0.04, 0.40	1.59	.11
Hispanic	-0.20	-0.46, 0.06	1.49	.14
African American	0.13	-0.23, 0.50	0.72	.47
High school graduate or higher	-0.04	-0.39, 0.30	0.25	.81
Student	0.15	-0.07, 0.37	1.34	.18
Married	-0.02	-0.20, 0.17	-0.16	.87
Living with spouse	-0.20	-0.50, 0.10	1.28	.20
Household size	-0.02	-0.07, 0.03	0.67	.51
Lived at residence >10-y	-0.12	-0.37, 0.13	0.93	.35

HRSD<sub>17</sub> indicates 17-item Hamilton Rating Scale for Depression; IVR, interactive voice response; QIDS-SR<sub>16</sub>, 16-item Quick Inventory of Depressive Symptomatology.

<sup>a</sup>Estimates were adjusted for covariates: patient age, Hispanic and African American status (patients could indicate both), sex, marital status (married vs not married for all reasons), education (high school graduate or higher vs less than high school graduate), private insurance status (yes/no), paid employment (full or part time vs unemployed for all reasons), baseline QIDS-SR<sub>16</sub> score, and order of IVR call where 1 indicates the first completed call, 2 indicates the second completed call, and so forth. If the selected predictor also was a covariate, the term was entered only once in the final model.

<sup>b</sup>For 6858 IVR calls among 2677 consenting patients for whom records were available at 39 sites. Reporting bias equals the number of depression-related visits computed from IVR responses minus visits computed from medical records for the same 3-month reporting period. Because patients on average underreported visits, a positive coefficient reflects less disparity, whereas a negative coefficient reflects greater disparity between IVR responses and medical records.

<sup>c</sup>IVR calls were initiated by respondents who were asked and who completed the scripted utilization-of-care questions.

<sup>d</sup>Symptom response defined by at least a 50% reduction in baseline scores on the QIDS-SR<sub>16</sub>—self-report at exit.

<sup>e</sup>Symptom remission defined by achieving a score of 5 or lower on the QIDS-SR<sub>16</sub>—self-report at exit.

<sup>f</sup>Symptom remission defined by a score of 7 or lower on the HRSD<sub>17</sub> at exit.

<sup>g</sup>Patients strongly agree or agree (rather than neutral, disagree, or strongly disagree) that if they get the help they need from a doctor, they will be better able to make important decisions affecting life, or they will be better able to enjoy things.

<sup>h</sup>Patients find the impact of family and friends on their health condition to be very helpful, moderately helpful, or minimally helpful (rather than neutral, minimally, moderately more, or much more difficult).

### Take-Away Points

The feasibility of interactive voice response (IVR) systems to collect use-of-care data was assessed in a large clinical trial (STAR\*D) involving 41 clinics and 4041 patients with major depression.

- Moderate intraclass correlation and underreporting biases were found when patient responses were compared with medical records.
- Reporting biases varied with baseline symptoms and IVR experience, but not treatment outcomes, demographic characteristics, or care attitudes.
- Clinical managers should use IVR systems to collect service histories only after patients are properly trained and responses monitored for consistency.

Underreporting biases present problems for clinicians evaluating the service histories of individual patients. However, these biases are stable across demographic (sex, race, ethnicity) and patient outcome groups where plan administrators may use IVR data to make between-group comparisons. That is, differences between groups will be unbiased whenever the under-report-

ing biases in each group cancelled out in the difference. Our findings contrast with those of Wallihan et al, who found that African Americans were more likely to understate ambulatory care visits during telephone surveys over a 1-year period,<sup>54</sup> but parallel those of Rozario et al, who found no racial effects.<sup>55</sup>

Response inconsistencies did vary by demographic characteristics. For instance, men made fewer inconsistent responses than women. This finding is consistent with the observations of Ritter et al,<sup>52</sup> Raina et al,<sup>53</sup> Wallihan et al,<sup>54</sup> and Rozario et al,<sup>55</sup> who found, with smaller sample sizes, that men had numerically but not statistically fewer discrepancies than women.

Patients who presented with more severe depressive symptoms also made fewer inconsistent responses and had fewer reporting biases. These findings are consistent with those of Raina et al,<sup>53</sup> who found smaller reporting biases when patients claimed poorer baseline health status. Small reporting biases leads to speculation that sicker patients, presumably in search of symptom relief, may have been better motivated to provide more accurate information.

There are several limitations to the study. The reliability of self-reported use of care was based on data from 2677 participants (from among the original 3789 patients who completed at least 1 use-of-care IVR call) who signed releases and from whom medical records were obtained. Differences between these samples are described elsewhere.<sup>1</sup> On the other hand, our estimates may hold external validity because patients were drawn from diverse clinical settings across the United States. Response inconsistencies may not necessarily reflect bad data, as patients may have reported zero volume in order to correct a prior “yes” response. Response biases were computed only for depression-related outpatient visits. Finally, this study compared responses from only 1 survey mode: IVR. Reporting differences have been observed between mail and Web-based surveys versus phone<sup>56</sup> and IVR<sup>57</sup> systems.

Decision makers planning to use IVR systems to collect information on patient service histories should consider the following. First, IVR systems should be designed to alert patients whenever their responses are inconsistent and allow

## DISCUSSION

Patients often seek care from diverse, off-network, and out-of-plan healthcare providers. Such information is important to clinicians preparing treatment plans, administrators monitoring healthcare costs, and scientific investigators conducting cost-outcome studies. However, access to off-plan medical records is complicated by HIPAA regulations and patients who often refuse to grant access to such records.<sup>1</sup> For STAR\*D, 94% of study enrollees (3789 of 4041) completed a use-of-care IVR call, while 77% (3116 of 4041) granted access, with only 66% (2677 of 4041) actually releasing records. In many cases, patient self-reports may be the only practical source of information decision makers have to determine use from all of the patient’s care providers.

This study evaluated a computer-scripted, IVR-formatted survey, the UAC-IVR, designed to collect self-reported use of health services. When compared with actual medical records, IVR responses tended to underreport depression-related outpatient visits by 20%, but with moderate agreement at 0.68 intraclass correlation. These findings, however, are comparable to other means of collecting self-reported health data, such as Van den Brink et al’s weekly/monthly diaries for healthcare products ( $r = 0.38$  to  $0.75$ )<sup>51</sup>; Ritter et al’s mailed questionnaire and 6-month reporting period for outpatient ( $r = 0.64$ ), emergency room ( $r = 0.60$ ), and inpatient care ( $r = 0.74$ )<sup>52</sup>; Raina et al’s telephone survey of seniors covering a 1-year observation interval for hospital care (intraclass correlation coefficient [ICC] = 0.50) and outpatient services (ICC = 0.25-0.41)<sup>53</sup>; and face-to-face structured interviews of veterans with mood disorders measuring outpatient encounters within 90-day intervals (ICC = 0.74).<sup>3</sup>

them to make corrections. The IVR systems also should record each occurrence, as inconsistent responses are related to reporting biases and may be symptomatic of other reporting problems. Second, response consistency improved and reporting errors eventually diminished as responders gained experience. Thus, several practice runs with feedback are highly recommended before using IVR data for decision-making purposes. Third, for accurate information, responders should be given reminders that the information being collected over the IVR call is important. Finally, IVR reporting errors must be weighted against inaccessible medical records, data collection efficiency, flexible response times for responders, and biased samples when plan enrollees are excluded from analyses because they were unwilling to share off-plan medical records.

**Author Affiliations:** From the Department of Psychiatry (TMK, MHT, AW, KS-W, AJR), University of Texas Southwestern, Dallas; the Veterans Health Administration (TMK), Washington, DC; the Department of Psychiatry (MF), Massachusetts General Hospital, Boston; Healthcare Technology Systems, Inc (JHG, JCM), Madison, WI; and the Department of Epidemiology (SRW), University of Pittsburgh, Pittsburgh, PA.

**Funding Source:** This project was funded in whole or in part with federal funds from the National Institute of Mental Health, National Institutes of Health, under contract N01-MH-90003 (A. J. Rush, MD, Principal Investigator). The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services or the Department of Veterans Affairs, nor does mention of trade names, commercial products, or organizations imply endorsement by the US government.

**Author Disclosure:** Dr Trivedi reports having received research support, having served as a consultant and having received lecture fees from several companies, including Bristol-Myers Squibb, Cephalon, Cyberonics, Eli Lilly, Forest, GlaxoSmithKline, Janssen, Organon, Pharmacia & Upjohn, Solvay, and Wyeth. Dr Fava reports having equity interests in Compellis and Med-Avante, and has received copyright royalties for the MGH CPFQ, DESS, and SAFER. Dr Fava also reports patent applications for SPCD and for a combination of azapirones and bupropion in major depressive disorders. Dr Fava has received research support, lecture fees, and consulting fees from several companies, including AstraZeneca, Boehringer-Ingelheim, Novartis, Pfizer, and Roche. Drs Greist and Mundt are employees of Healthcare Technology Systems, the company that provided the interactive voice response (IVR) system in this study. Dr Greist has received research support, lecture fees, and consulting fees from several companies, including Bristol-Myers Squibb, Cyberonics, Eli Lilly, GlaxoSmithKline, Ortho-McNeil, Pfizer, and Solvay. Dr Mundt reports owning stock in Healthcare Technology Systems. Dr Rush reports receiving royalties from Healthcare Technology Systems for the IVR version of the Quick Inventory of Depressive Symptomatology (QIDS-SR). Dr Wisniewski reports having served as a consultant to Cyberonics, Bristol-Myers Squibb, ImaRx Therapeutics, and Organon. The other authors (TMK, AW, KS-W) report no relationship or financial interest with any entity that would pose a conflict of interest with the subject matter of this article.

**Authorship Information:** Concept and design (TMK, MHT, MF, JCM, KS-W, AJR, SRW); acquisition of data (TMK, AW, MF, JHG, JCM, AJR); analysis and interpretation of data (TMK, MHT, AW, JHG, KS-W, AJR); drafting of the manuscript (TMK, JCM, AJR); critical revision of the manuscript for important intellectual content (TMK, MHT, AW, MF, JHG, JCM, KS-W, AJR, SRW); statistical analysis (TMK); provision of study materials or patients (JCM); obtaining funding (AJR, SRW); administrative, technical, or logistic support (AW, JHG, JCM, AJR); and supervision (AW).

**Address correspondence to:** T. Michael Kashner, PhD, JD, MPH, Department of Psychiatry, University of Texas Southwestern, 5323 Harry Hines Blvd, Dallas, TX 75390-9086. E-mail: michael.kashner@utsouthwestern.edu.

## REFERENCES

1. Kashner TM, Trivedi MH, Wicker A, et al. Consent bias in accessing medical records in clinical trials: a STAR\*D report. *International Journal of Methods in Psychiatric Research*. In press.
2. Kashner TM. Agreement between administrative files and written medical records: a case of the Department of Veterans Affairs. *Med Care*. 1998;36(9):1324-1336.
3. Kashner TM, Stensland MD, Lind L, et al. Measuring use and cost of care for patients with mood disorders: the utilization and cost inventory. *Med Care*. 2009;47(2):184-190.
4. Rush AJ, Fava M, Wisniewski SR, et al. Sequenced treatment alternatives to relieve depression (STAR\*D): rationale and design. *Control Clin Trials*. 2004;25(1):119-142.
5. Fava M, Rush AJ, Trivedi MH, et al. Background and rationale for the sequenced treatment alternatives to relieve depression (STAR\*D) study. *Psychiatr Clin North Am*. 2003;26(2):457-494.
6. Trivedi MH, Rush AJ, Wisniewski SR, et al. Evaluation of outcomes with citalopram for depression using measurement-based care in STAR\*D: implications for clinical practice. *Am J Psychiatry*. 2006;163(1):28-40.
7. Rush AJ, Trivedi MH, Wisniewski SR, et al. Bupropion-SR, sertraline, or venlafaxine-XR after failure of SSRIs for depression. *N Engl J Med*. 2006;354(12):1231-1242.
8. Trivedi MH, Fava M, Wisniewski SR, et al. Medication augmentation after the failure of SSRIs for depression. *N Engl J Med*. 2006;354(12):1243-1252.
9. Schumacher RM, Hardzinski ML, Schwartz AL. Increasing the usability of interactive voice response systems: research and guidelines for phone-based interfaces. *Hum Factors*. 1995;37:251-264.
10. Ewing H, Reesal R, Kobak KA, Greist JH. Patient satisfaction with computerized assessment in a multicenter clinical trial. Paper presented at: 38th Annual Meeting of the New Clinical Drug Evaluation Unit (NCDEU); June 10-13, 1998; Boca Raton, FL.
11. Katzelnick DJ, Kobak KA, Helstad CP, et al. The direct and indirect costs of social phobia in managed care patients. Paper presented at: 37th Annual Meeting of the American College of Neuropsychopharmacology; December 14-18, 1998; Las Croabas, Puerto Rico.
12. Agel J, Rockwood T, Mundt JC, Greist JH, Swiontkowski M. Comparison of interactive voice response and written self-administered patient surveys for clinical research. *Orthopedics*. 2001;24(12):1155-1157.
13. Greist JH, Klein MH, Van Cura LJ. A computer interview for psychiatric patient target symptoms. *Arch Gen Psychiatry*. 1973;29(2):248-253.
14. Richards JS, Fine RR, Wilon TL, Rogers JT. A voice-operated method for administering the MMPI. *J Pers Assess*. 1983;47(2):167-170.
15. Mundt JC. Interactive voice response systems in clinical research and treatment. *Psychiatr Serv*. 1997;48(5):611-612.
16. Mundt JC, Bohn MJ, King M, Hartley MT. Automating standard alcohol use assessment instruments via interactive voice response technology. *Alcohol Clin Exp Res*. 2002;26(2):207-211.
17. Lee H, Friedman ME, Cukor P, Ahem D. Interactive voice response system (IVRS) in health care services. *Nurs Outlook*. 2003;51(6):277-283.
18. Searles JS, Perrine MW, Mundt JC, Helzer JE. Self-report of drinking by touch-tone telephone: extending the limits of reliable daily contact. *J Stud Alcohol*. 1995;56(4):375-382.
19. Nakagawa A, Marks IM, Park JM, Bachofen M, Baer L, Dotti SL. Self-treatment of obsessive-compulsive disorder guided by manual and computer-conducted telephone interview. *J Telemed Telecare*. 2000;6(1):22-26.
20. Bardone AM, Krahn DD, Goodman BM, Searles JS. Using interactive voice response technology and timeline follow-back methodology in studying binge eating and drinking behavior: different answers to different forms of the same question? *Addict Behav*. 2000;25(1):1-11.
21. Greist JH, Klein MH, Erdman HP. Comparison of computer- and interview-administered versions of the Diagnostic Interview Schedule. *Hosp Community Psychiatry*. 1987;38(12):1304-1311.
22. Mundt JC, Searles JS, Perrine MW, Walter D. Conducting longitudinal studies of behavior using interactive voice response technology. *Int J Speech Technol*. 1997;2:21-31.
23. Aiernagno SA, Cochran D, Feucht TE, Stephens RC, Butts JM, Wolfe SA. Assessing substance abuse treatment needs among the

- homeless: a telephone-based interactive voice response system. *Am J Public Health*. 1996;86(11):1626-1628.
24. **Mundt JC, Ferber KL, Rizzo M, Greist JH.** Computer automated dementia screening using a touch-tone telephone. *Arch Intern Med*. 2001;161(20):2481-2487.
25. **Mundt JC, Clarke GN, Burroughs D, Brennehan DO, Greist JH.** Effectiveness of antidepressant pharmacotherapy: the impact of medication compliance and patient education. *Depression Anxiety*. 2001;13(1):1-10.
26. **Krystal AD, Walsh JK, Laska E, et al.** Sustained efficacy of eszopiclone over 6 months of nightly treatment: results of a randomized, double-blind, placebo-controlled study in adults with chronic insomnia. *Sleep*. 2003;26(7):793-799.
27. **Schneider SJ, Schwartz MD, Fast J.** Computerized, telephone-based health promotion, I: smoking cessation program. *Computers in Human Behavior*. 1995;11:135-148.
28. **Kobak KA, Mundt JC, Greist JH, Katzelnick DJ, Jeferson JW.** Computer assessment of depression: automating the Hamilton Depression Rating Scale. *Drug Info J*. 2000;34:145-156.
29. **Greist JH, Marks IM, Baer L, et al.** Self-treatment for obsessive compulsive disorder using a manual and a computerized telephone interview: A U.S.-U.K. study. *MD Comput*. 1998;15(3):149-157.
30. **Piette JD.** Interactive voice response systems in the diagnosis and management of chronic disease. *Am J Manag Care*. 2000;6(7):817-827.
31. **Spitzer RL, Williams JB, Kroenke K.** Utility of a new procedure for diagnosing mental disorders in primary care: the PRIME-MD 1000 study. *JAMA*. 1994;272(22):1749-1756.
32. **Kobak KA, Taylor LV, Dotti SL.** A computer-administered telephone interview to identify mental disorders. *JAMA*. 1997;278(11):905-910.
33. **Kashner TM, Suppes T, Rush AJ, Altshuler KZ.** Measuring use of outpatient care among mentally ill individuals: a comparison of self-reports and provider records. *Eval Program Plann*. 1999;22:31-39.
34. **Kashner TM, Rush AJ, Altshuler KZ.** Measuring costs of guideline-driven mental health care: the Texas Medication Algorithm Project. *J Ment Health Policy Econ*. 1999;2(3):111-121.
35. **Rush AJ, Crismon ML, Kashner TM, et al.** Texas Medication Algorithm Project, phase 3 (TMAP-3): rationale and study design. *J Clin Psychiatry*. 2003;64(4):357-369.
36. **Kashner TM, Rush AJ, Crismon ML, et al.** An empirical analyses of cost outcomes of the Texas Medication Algorithm Project. *Psychiatr Serv*. 2006;57(5):648-659.
37. **American Medical Association.** *Current Procedural Terminology*. Reston, VA: St. Anthony; 2004.
38. **Centers for Medicare & Medicaid Services.** Healthcare Common Procedure Coding System (HCPCS). <http://www.cms.hhs.gov/MedHCPCSGenInfo/>. Accessed March 6, 2009.
39. **American Psychiatric Association.** *Diagnostic and Statistical Manual of Mental Disorders*. 4th ed. Washington, DC: American Psychiatric Association; 1994.
40. **Ware J Jr, Kolinsky M, Keller SD.** The 12-item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. *Med Care*. 1996;34(3):220-223.
41. **Hamilton M.** A rating scale for depression. *J Neurol Neurosurg Psychiatry*. 1960;23:56-62.
42. **Hamilton M.** Development of a rating scale for primary depressive illness. *Br J Soc Clin Psychol*. 1967;6(4):278-296.
43. **Rush AJ, Carmody TJ, Ibrahim HM, et al.** Comparison of self-report and clinician ratings on two inventories of depressive symptomatology. *Psychiatr Serv*. 2006;57(6):829-837.
44. **Trivedi MH, Rush AJ, Ibrahim HM, et al.** The Inventory of Depressive Symptomatology, Clinician Rating (IDS-C) and Self-Report (IDS-SR), and the Quick Inventory of Depressive Symptomatology, Clinician Rating (QIDS-C) and Self-Report (QIDS-SR) in public sector patients with mood disorders: a psychometric evaluation. *Psychol Med*. 2004;34(1):73-82.
45. **Rush AJ, Trivedi MH, Ibrahim HM, et al.** The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), Clinician Rating (QIDS-C), and Self-Report (QIDS-SR): a psychometric evaluation in patients with chronic major depression [Published correction appears in: *Biol Psychiatry*. 2003;54(5):585]. *Biol Psychiatry*. 2003;54(5):573-583.
46. **Rush AJ, Bernstein IH, Trivedi MH, et al.** An evaluation of the Quick Inventory of Depressive Symptomatology and the Hamilton Rating Scale for Depression: a Sequenced Treatment Alternatives to Relieve Depression trial report. *Biol Psychiatry*. 2006;59(6):493-501.
47. **Shrout PE, Fleiss JL.** Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420-428.
48. **Fleiss JL.** *The Design and Analysis of Clinical Experiments*. New York: John Wiley; 1986.
49. **Bryk AS, Raudenbush SW.** *Hierarchical Linear Models*. Newbury Park, CA: Sage; 1992.
50. **Andersen RM.** Revisiting the behavioral model and access to medical care: does it matter? *J Health Soc Behav*. 1995;36(1):1-10.
51. **van den Brink M, van den Hout WB, Stiggelbout AM, van de Velde CJH, Kievit J.** Cost measurement in economic evaluations of health care: whom to ask? *Med Care*. 2004;42(8):740-746.
52. **Ritter PL, Stewart AL, Kaymaz H, Sobel DS, Block DA, Lorig KR.** Self-reports of health care utilization compared to provider records. *J Clin Epidemiol*. 2001;54(2):136-141.
53. **Raina P, Torrance-Rynard V, Wong M, Woodward C.** Agreement between self-reported and routinely collected health-care utilization data among seniors. *Health Serv Res*. 2002;37(3):751-774.
54. **Wallihan DB, Stump TE, Callahan CM.** Accuracy of self-reported health services use and patterns of care among urban older adults. *Med Care*. 1999;37(7):662-670.
55. **Rozario PA, Morrow-Howell N, Proctor E.** Comparing the congruency of self-report and provider records of depressed elders' service us by provider type. *Med Care*. 2004;42(10):952-959.
56. **Brewer NT, Hallman KW, Fiedler N, Kipen HM.** Who do people report better health by phone than by mail? *Med Care*. 2004;42(9):875-883.
57. **Rodríguez HP, von Glahn T, Rogers WH, Chang H, Fanjiang G, Safran DG.** Evaluating patients' experiences with individual physicians: a randomized trial of mail, Internet, and interactive voice response telephone administration of surveys. *Med Care*. 2006;44(2):167-174. ■