Test–retest reliability in a computer-based medical history

Warner V Slack,[1] Hollis B Kowaloff,[1] Roger B Davis,[1,2] Tom Delbanco,[2] Steven E Locke,[1] Howard L Bleich[1]

► Additional figures, table and appendix are published online only. To view these files please visit the journal online (www.jamia.org).

[1]Division of Clinical Informatics, Department of Medicine, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, Massachusetts, USA
[2]General Medicine and Primary Care, Department of Medicine, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, Massachusetts, USA

Correspondence to
Dr Warner V Slack, Division of Clinical Informatics, Department of Medicine, Beth Israel Deaconess Medical Center and Harvard Medical School, 1330 Beacon Street, Brookline, MA 024446, USA;
wslack@bidmc.harvard.edu

Received 14 May 2010
Accepted 3 November 2010
Published Online First
27 November 2010

ABSTRACT

The authors developed a computer-based medical history for patients to take in their homes via the internet. The history consists of 232 'primary' questions asked of all patients, together with more than 6000 questions, explanations, and suggestions that are available for presentation as determined by a patient's responses. The purpose of this research was to measure the test–retest reliability of the 215 primary questions that have preformatted, mutually exclusive responses of 'Yes,' 'No,' 'Uncertain (Don't know, Maybe),' 'Don't understand,' and 'I'd rather not answer.' From randomly selected patients of doctors affiliated with Beth Israel Deaconess Medical Center in Boston, 48 patients took the history twice with intervals between sessions ranging from 1 to 35 days (mean 7 days; median 5 days). High levels of test–retest reliability were found for most of the questions, but as a result of this study the authors revised five questions. They recommend that structured medical history questions that will be asked of many patients be measured for test–retest reliability before they are put into widespread clinical practice.

INTRODUCTION

We developed a computer-based medical history to be taken by patients in their home by means of the internet and then studied the test–retest reliability of the 215 questions in the interview that were asked of all patients, were of fixed wording, and had an identical set of possible responses.

BACKGROUND

The medical history has received little scientific attention compared with the other components of the diagnostic process. Advances in diagnostic imaging and analytical chemistry have far outpaced advances in our knowledge about the medical history.

The reasons are understandable. Fortuitous circumstances can influence both the doctor as interviewer and the patient as respondent, and the doctor, motivated to confirm clinical judgments, can unwittingly bias questions and elicit responses that conform to expectations.[1–3] These hard-to-control factors have motivated investigators to turn to the study of structured questions, with one wording and a limited number of possible responses.[4 5]

The validity of a medical history question—the degree to which it elicits the information it is intended to elicit—is dependent in part on the degree to which it elicits identical responses when asked a second time.[6] It can be argued, therefore, that structured medical history questions should be studied for their test–retest reliability before they are incorporated into clinical practice.

In an early study,[7] Collen and his colleagues measured the test–retest reliability of 204 general medical history questions printed on Hollerith cards and presented to patients twice with an interval of 30 minutes between presentations. Patients responded to each question by placing its card in one of two boxes labeled 'yes' or 'no.' The cards from each box for each presentation were then tabulated by a computer. In their study, 95% of the patients changed fewer than 6% of their responses; patients were more likely to switch from 'yes' to 'no' than from 'no' to 'yes;' and questions that were 'long, complex, or vague' were least reliable. More recent studies have focused on the reliability of questions presented on paper-based, self-administered questionnaires pertaining to specific medical problems.[8–13]

The digital computer, programmed to interact directly with the patient—to engage in meaningful dialog and to explore medical problems in detail—has also been used to conduct structured medical interviews[14 15]; and studies over the years have demonstrated the potential of patient–computer dialog to obtain comprehensive, accurate medical histories that are well received by patients and doctors,[16–22] to obtain responses to structured questions that are comparable to those of a clinician interviewer,[3 23] and to obtain sensitive, potentially embarrassing information that might not otherwise have been obtained by the clinician interviewer.[24–30]

Although once restricted to experimental settings, research with the computer as an interviewer can now be performed with programs delivered to the patient's home over the internet[21 22] and that can also be used to test the reliability of the questions in the interviews.[31] We have developed a detailed, interactive, computer-based medical history to be taken by patients in the privacy of their homes, prior to their first appointment with a primary care doctor. In preparation for a clinical trial with this interview, we studied the test–retest reliability of 215 of its questions.

Methods

The general medical history, developed on the basis of our experience over the years,[14 16–19 22–24] contains 232 'primary' questions, which are asked of all patients who take the history, together with more than 6000 qualifying questions (with varying sets of possible responses), explanations,

and suggestions available for presentation as determined by the patient's responses and the branching logic of the program. We have limited the results in this report to our analysis of the 215 primary questions that have the same preformatted, mutually exclusive responses of 'Yes,' 'No,' 'Uncertain (Don't know, Maybe),' 'Don't understand,' and 'I'd rather not answer.' (We have not included the remaining 17 primary questions, which have varying sets of possible responses.)

The interview is divided into 24 sections including family history, social history and a review of systems. A summary of the history is generated for the doctor and patient to review at the time of the patient's visit (see figure 1, available as an online data supplement at www.jamia.org).

From the patients of 11 primary care doctors affiliated with Boston's Beth Israel Deaconess Medical Center, we randomly selected 902 patients who had access to the internet in their homes. These patients were then apprised of the study by an email message that pointed them to an online description and a request for informed consent. Patients were advised that their responses would be used solely to evaluate the interview and would play no role in their medical care.

An automatically generated electronic mail message asked the patients to take the interview the first time, and 3 days after they had taken it, a second message asked them to take it again. Subsequently if needed, a reminder was sent at 3-day intervals. Each patient who completed the study was paid US$50.

During the second interview, when a response to a question under study differed from that made during the first interview, the patient was asked to choose from four reasons: 'My medical situation changed,' 'I clicked on the wrong choice,' 'I'm not really sure about the answer,' and 'I didn't understand the question,' or to bypass these if none pertained. The patient was then asked the question a third time as an opportunity to resolve the inconsistency.

For our analyses, we computed the percentage of agreement for each of the questions, together with Cohen's Kappa Index of Reliability.[32] Influenced by the criteria for the κ suggested by Landis and Koch[33]—excellent agreement (>0.75) and good to fair agreement (0.75–0.40)—we selected as our indicators of acceptable reliability, either κ values >0.60 or percentages of agreement >90. We used McNemar's test[34] to examine whether inconsistent responses tended to be in a particular direction. In our computations, we collapsed responses of 'Uncertain (Don't know, Maybe),' 'Don't understand,' and 'I'd rather not answer' into a single 'uncertainty' group.

To assess whether the likelihood of inconsistent responses was associated with either the patient's age or the time interval between the two interviews, we fitted logistic regression models. We used generalized estimating equation methods[35] to account for within-patient correlation. We report ORs and 95% CIs based on these models, as well as Wald tests comparing categories of age and time interval.

RESULTS

Of the 902 patients approached, 48 (37 women and 11 men between the ages of 20 and 77 years) took the entire interview (including branch points) twice, with intervals between sessions ranging from 1 to 35 days (mean 7 days; median 5 days) (see online supplementary figure 2). The 48 volunteers who answered the 215 questions during both interviews provided 10 320 responses, of which 37 from one patient were inadvertently lost, yielding 10 283 responses available for study. Of these, 9617 (93%) were identical with responses to the same questions during the first interview, and 666 were different. There was >90% agreement for 155 of the 215 questions (including 60 at 100% agreement). The κ scores were >0.75 for 96 of the questions (including 36 at 1.0) and >0.60 for 142 of the questions (our criterion for good agreement). Twenty-four of the questions had no κ scores, indicating identical responses—all of which were 'no'—by all 48 patients upon both asking and re-asking.

Of the 215 questions, 132 had both percentages of agreement >90 and κ scores either >0.60 or with no score, and an additional 57 questions had either percentages of agreement >90 or κ scores >0.60. We deemed these 189 questions (88%) to be sufficiently reliable within their clinical context to remain in the interview unrevised. The remaining 26 questions, which had both percentages of agreement <90 and κ scores <0.60, were responsible for 205 (31%) of the 666 inconsistencies, and we selected these for further evaluation and possible rewording.

Of the 26 questions with poor reliability, 18 asked the patients to recall from memory any additional problems not specifically addressed in the interview, and, upon careful examination, we could not think of a way to improve them. Of the remaining eight questions with poor reliability, three asked for information that may be unavailable or difficult to recall—family history of yellow jaundice, anemia, and abnormal bleeding—and we considered these questions to be acceptable as well.

In an effort to improve their reliability, we revised the remaining five questions (see appendix 1, available online). From among these, 'Have you been having trouble sleeping at night?' was split into two questions—'Do you have trouble falling asleep at night?' and 'If you wake up, do you have trouble falling back to sleep?'

Patients were more likely to respond consistently when their first response was 'no' than when it was either 'yes' ($p<0.0001$, McNemar's test) or 'uncertain' ($p<0.0001$). When inconsistent, patients were more likely to switch from 'yes' to 'no' than from 'no' to 'yes' ($p<0.0001$). When the first response was 'yes' (N=300), in 62% of the cases the patient returned to 'yes' when asked the third time. In contrast, when the first response was 'no' (N=185), in only 47% of the cases did the patient return to 'no' when asked a third time. When their first response was 'uncertain,' patients were more likely to switch to 'no' than to 'yes' upon the second asking ($p<0.0001$), but when patients veered from their first response of 'uncertain' (N=181), in only 45% of the cases did they return to 'uncertain' upon the third asking (online supplementary table 1).

Patients attributed 159 of their inconsistencies (24%) to 'clicked on the wrong choice,' and with these they returned to their first answer 87 times (55%) when asked the third time. Patients attributed 151 of their inconsistencies (23%) to 'not really sure about the answer' and returned to their first answer 38 times (25%) when asked the third time. Patients attributed 40 of their inconsistencies (6%), representing 25 questions, to 'medical situation changed' and stayed with their second response 38 times (95%) when asked the third time. In three instances of inconsistency, patients explained with 'didn't understand the question,' and, in 313 instances (47%), they gave no reason.

In unadjusted logistic regression models, we found that both patient age ($p=0.0009$) and the time interval between interviews ($p=0.01$) were associated with the proportion of inconsistent responses. However, when we included both factors in the model simultaneously, only age was significantly associated with inconsistency. In particular, patients over 40 were more likely to have inconsistent responses. Further studies will be

required to determine if age is responsible for inconsistency, or if it is a marker for more illness and therefore a marker for more details (online supplementary table 2).

Patients used the 'uncertainty' options sparingly. During the first session, four patients used 'Don't understand' one time—two patients in response to the same question about nasal discharge, which turned out to be one of the five questions with poor reliability that we subsequently revised. Patients used 'Uncertain (Don't know, Maybe)' 453 times, and in two instances invoked 'I'd rather not answer.'

With the first interview session, the 48 patients were presented with a mean of 545 question-screens and took between 45 and 90 minutes to complete the interview (based on an estimated 7 s per screen[36]); and patients were favorable in their assessment of the interview when asked a concluding set of questions (online supplementary figure 3).

DISCUSSION

In our study, 'no' was a more reliable first response than 'yes,' and, as with Collen *et al*'s finding,[7] patients were more likely to change from 'yes' to 'no' than from 'no' to 'yes.' Perhaps patients tend to err on the side of affirmative responses so as not to overlook a medical problem that might be important. Still, when their initial response was 'yes,' patients more often than not returned to 'yes' upon the third asking.

When inconsistencies did occur, we found the request for an explanation and the use of a third, 'tie-breaker' question helpful when evaluating the responses. In 38 of the 40 inconsistencies attributed by patients to medical changes, the patient's response to the third asking was consistent with the response to the second asking. We could deduce therefore that the medical change was real, and that the questions as presented had been reliable. Furthermore, if consistency between the second and third asking is an indication of what might be called a secondary reliability, the tie breaker can also help to compare 'false-positive' and 'false-negative' responses. Discounting the 38 changes in responses attributed to medical changes, we counted relatively more 'true' false-positive response sequences of 'yes' to 'no' to 'no' (41 or 2% of the 1929 first responses of 'yes') than false-negative sequences of 'no' to 'yes' to 'yes' (40 or 0.4% of the 8028 first responses of 'no').

The good reliability of 189 (88%) of the 215 questions in the study is reassuring. Of the 26 questions with poor reliability, 18 were of a similar construct: they asked patients to consider from memory a broad range of possible medical problems before they could respond. The most frequent explanation for the inconsistency with these 18 questions was 'not really sure about the answer.'

When test–retest reliability is measured, the time interval between 'askings' is important. If the interval is too short, the patient may respond consistently by way of rote memory. If too long, the clinical situation may have changed. Streiner and Norman suggest an optimal interval of between 2 and 14 days.[6] Although five of our intervals were longer than 14 days, when we included both age and time interval in our regression model, only age—with patients over 40 years old—was associated with increased inconsistency in responses. It is possible that the older patients were less comfortable with the computer, or that older patients are inherently less likely to give consistent responses. On the other hand, older patients are more likely to have medical problems that complicate recall.

Prior to the start of this study, we did our best to perfect each question. We asked 10 patients to read each question aloud and to explain the meaning. We then revised those questions that were not readily understood. Only after we thought the questions were as good as we could make them did we begin this study. Still, the study enabled us to detect five additional questions that we considered to be in need of revision.

We conclude that test–retest reliability is an important and useful method for screening a structured interview for questions in need of revision. We recommend that structured medical history questions, whether presented by computer, paper questionnaire, or interpersonal dialogue, be routinely measured for test–retest reliability before being put into clinical practice.

Acknowledgments We are indebted to the participating patients and doctors and thank them for their help with this study.

Funding This study was supported in part by a grant from the National Library of Medicine (1 R01 LM008255-01A1) and by a grant from the Rx Foundation.

Competing interests WS is on the Scientific Advisory Board of the Eliza Corporation with stock options in the company. TD is a director of the Eliza Corporation and holds stock options in the company. SEL is a principal in Veritas Health Solutions LLC, a principal in Cognitive Behavioral Technologies LLC, an owner of Veritas Health Associates LLC, and consults for LifeOptions Group and Mensante Corp.

Ethics approval This study was conducted with the approval of the Beth Israel Deaconess Medical Center, Boston, Massachusetts.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Kahn RL,** Cannell CF. *The Dynamics of Interviewing*. New York (NY): John Wiley and Sons, 1957.
2. **Duncan S,** Finkelstein MJ. Nonverbal communications of experimental bias. *Proc Annu Conv Amer Psychol Assoc* 1969;**4**:369–70.
3. **Slack WV,** Slack CW. Talking to a computer about emotional problems: a comparative study. *Psychotherapy: Theory, Research, and Practice* 1977;**14**:156–64.
4. **Cochrane AL,** Chapman PJ, Oldham PD. Observer's errors in taking medical histories. *Lancet* 1951;**1**:1007–9.
5. **Brodman K,** Erdmann AJ, Lorge I, *et al*. Cornell Medical Index: adjunct to medical interview. *JAMA* 1949;**140**:530–4.
6. **Streiner DL,** Norman GR. *Health measurement scales: a practical guide to their development and use*. 2nd edn. Oxford: Oxford University Press, 1995.
7. **Collen MF,** Cutler JL, Siegelaub MA, *et al*. Reliability of a self-administered medical questionnaire. *Arch Intern Med* 1969;**123**:664–71.
8. **Ford JJ,** Story I, McMeeken J. The test-retest reliability and concurrent validity of the Subjective Complaints Questionnaire for low back pain. *Man Ther* 2009;**14**:283–91.
9. **Allison C,** Williams J, Scott F, *et al*. The Childhood Asperger Syndrome Test (CAST): test-retest reliability in a high scoring sample. *Autism* 2007;**11**:173–85.
10. **Thiel A,** Paul T. Test-retest reliability of the eating disorder inventory 2. *J Psychosom Res* 2006;**61**:567–9.
11. **Bradbury BD,** Brooks DR, Brawarsky P, *et al*. Test-retest reliability of colorectal testing questions on the Massachusetts Behavioral Risk Factor Surveillance System (BRFSS). *Prev Med* 2005;**41**:303–11.
12. **Posternak MA,** Young D, Sheeran T, *et al*. Assessing past treatment history: test-retest reliability of the treatment response to antidepressant questionnaire. *J Nerv Ment Dis* 2004;**192**:98–102.
13. **Reider CR,** Hubble JP. Test-retest reliability of an epidemiological instrument for Parkinson's disease. *J Clin Epidemiol* 2000;**53**:863–5.
14. **Slack WV,** Hicks GP, Reed CE, *et al*. A computer-based medical history system. *N Engl J Med* 1966;**274**:194–8.
15. **Mayne JG,** Weksel W, Sholtz PN. Toward automating the medical history. *Mayo Clin Proc* 1968;**43**:1–25.
16. **Slack WV.** A history of computerized medical interviews. *MD Comput* 1984;**1**:52–9.
17. **Slack WV.** Patient-computer dialogue: a review. In: van Bemmel JH, McCray AT, eds. *Yearbook of Medical Informatics 2000: Patient-Centered Systems*. Stuttgart, (Germany): Schattauer, 2000:71–8.
18. **Slack WV.** Cybermedicine as a patient's assistant. In: Slack WV**.** *Cybermedicine: How Computing Empowers Doctors and Patients for Better Health Care*. Rev edn. San Francisco (Calif): Jossey-Bass, 2001:38–43.
19. **Slack WV.** A 67-year old man who e-mails his physician. *JAMA* 2004;**292**:2255–61.
20. **Bachman JW.** The patient-computer interview: a neglected tool that can aid the clinician. *Mayo Clin Pro* 2003;**78**:67–78.
21. **Adamson SC,** Bachman JW. Pilot study of providing online care in a primary care setting. *Mayo Clin Proc* 2010;**85**:704–9.

22. **Slack WV.** Patient-computer dialogue: a hope for the future. *Mayo Clin Proc* 2010;**85**:701—3.
23. **Slack WV,** Slack CW. Patient-computer dialogue. *N Engl J Med* 1972;**286**:1304—9.
24. **Slack WV,** Van Cura LJ. Patient reaction to computer-based medical interviewing. *Comput Biomed Res* 1968;**1**:527—31.
25. **Greist JH,** Gustafson DH, Stauss FF, *et al*. A computer interview for suicide-risk prediction. *Am J Psychiatry* 1973;**130**:1327—32.
26. **Lucas RW,** Card WI, Knill-Jones RP, *et al*. Computer interrogation of patients. *Br Med J* 1976;**2**:623—5.
27. **Locke SE,** Kowaloff HB, Hoff RG, *et al*. Computer-based interview for screening blood donors for risk of HIV transmission. *JAMA* 1992;**268**:1301—5.
28. **Turner CF,** Ku L, Rogers SM, *et al*. Adolescent sexual behavior, drug use, and violence: Increased reporting with computer survey technology. *Science* 1998;**280**:867—73.
29. **Kurth AE,** Martin DP, Golden MR, *et al*. A comparison between audio computer-assisted self-interviews and clinician interviews for obtaining the sexual history. *Sex Transm Dis* 2004;**31**:719—26.
30. **Mackenzie SL,** Kurth AE, Spielberg F, *et al*. Patient and staff perspectives on the use of a computer counseling tool for HIV and sexually transmitted infection risk reduction. *J Adolesc Health* 2007;**40**:572.e9—16.
31. **Brigham J,** Lessov-Schlaggaar CN, Javitz HS, *et al*. Test-retest reliability of web-based retrospective self-report of tobacco exposure and risk. *J Med Internet Res* 2009;**11**:e35.
32. **Cohen J.** Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968;**70**:213—20.
33. **Landis JR,** Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159—74.
34. **McNemar Q.** Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947;**12**:153—7.
35. **Zeger SL,** Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986;**42**:121—30.
36. **Slack WV,** Leviton A, Bennett SE, *et al*. Relation between age, education, and time to respond to questions in a computer-based medical interview. *Comput Biomed Res* 1988;**21**:78—4.