Agreement Between Patient and Proxy Assessments of Health-Related Quality of Life After Stroke Using the EQ-5D and Health Utilities Index

A. Simon Pickard, PhD; Jeffrey A. Johnson, PhD; David H. Feeny, PhD; Ashfaq Shuaib, MD; K.C. Carriere, PhD; Abdul M. Nasser, MD

Background and Purpose—Proxy informants can provide information on patients who are limited in ability to self-assess health-related quality of life (HRQL) after stroke. One alternative is to exclude assessments of such patients and attenuate generalizability. The purpose of this study was to examine patient-proxy agreement on the domains and summary scores of the EQ-5D and Health Utilities Index Mark 3 (HUI3) after stroke.

Methods—An observational longitudinal cohort of 124 patients hospitalized after ischemic stroke and their family caregivers completed the HRQL measures at baseline and were followed up for 6 months. Patient and proxy agreement was assessed by use of weighted κ or the intraclass correlation coefficient (ICC).

Results—At baseline, the more observable domains of HRQL demonstrated greater agreement than the more subjective components. Cross-sectional point estimates of agreement were generally acceptable (ICC 0.70) for the EQ-5D Index and HUI3 summary scores when assessed ≥1 month after baseline. Agreement between change scores was generally poor to fair (ICC <0.60), but systematic bias was not observed for the indirect preference-based summary scores between baseline and 6 months.

Conclusions—Results suggest that proxy assessments obtained 6 months after stroke are more reliable than those obtained within 2 to 3 weeks after stroke. Although proxy-assessed change scores for indirect preference-based summary scores of the EQ-5D and HUI3 provided suboptimal agreement with patient assessment, limited systematic bias may support their consideration as alternatives to missing data or statistical imputation. Further research into the validity and reliability of proxy assessments is suggested.

Key Words: observer variation ■ outcome ■ quality of life ■ stroke assessment

Stoke is often a debilitating condition that potentially affects many aspects of health-related quality of life (HRQL). Generic preference-based HRQL measures can be used to summarize a diverse range of stroke outcomes, to integrate the effects of mortality and morbidity, to compare the burden of stroke to other conditions, and to compare programs through cost utility analyses.1 However, the ability to self-assess HRQL after stroke can be limited by cognitive impairment, fatigue, stress, or communication problems.2 Patients who have receptive aphasia, for example, may be excluded from the patient-reported outcomes component of a clinical trial because they cannot respond on their own behalf initially after stroke but may be able to do so 6 months later. Thus, enlisting a family member to respond on behalf of the patient avoids systematic exclusion of such patients that may threaten the generalizability of study results. In addition, proxy respondents can reduce missing data, which can compromise power to detect changes over time or to detect differences between groups and creates potentially biased estimates as a result of nonrandomly missing data.3

The reliability of proxy raters has been examined in several independent studies of generic HRQL measures in stroke, including investigations of the Health Utilities Index Mark 2 (HUI2) and Mark 3 (HUI3),4 EQ-5D, Health Status Questionnaire,6 and Sickness Impact Profile.2 Consistent with the general literature on proxy respondents, these studies found that patient-proxy agreement was stronger for physically based, observable attributes than for less observable, psychosocial attributes. Recommendations varied, with investigators supportive of using proxies (Sickness Impact Profile, HUI2/3),2,4 not supportive (Health Status Questionnaire),6 or conditionally endorsing proxy respondents as reliable assessors of the more observable domains of HRQL (EQ-5D).5

Received June 12, 2003; final revision received September 4, 2003; accepted October 14, 2003.

From the College of Pharmacy, University of Illinois at Chicago (A.S.P.); Division of Neurology (A.M.N., A.S.), Faculty of Medicine and Dentistry (J.A.J., D.H.F.), Faculty of Pharmacy and Pharmaceutical Sciences (D.H.F.), and Department of Mathematics and Statistics (K.C.C.), University of Alberta, Edmonton, Alberta; Institute of Health Economics (J.A.J., D.H.F.), Edmonton, Alberta; and Health Utilities Incorporated (D.H.F.), Dundas, Ontario, Canada.

Reprint requests to A. Simon Pickard, PhD, College of Pharmacy, Room 164, 833 S Wood St (MC886), University of Illinois at Chicago, Chicago, IL 60612. E-mail: pickard1@uic.edu

© 2004 American Heart Association, Inc.

Stroke is available at http://www.strokeaha.org

DOI: 10.1161/01.STR.0000110984.91157.BD

by guest on April 13, 2017 http://stroke.ahajournals.org/ Downloaded from
We were interested in using a longitudinal design to examine the reliability of proxy assessments using both the EQ-5D and HUI3 for several reasons. Previous studies in stroke have investigated each measure independently using cross-sectional designs. A longitudinal study using several HRQL measures concomitantly would be useful to corroborate agreement on different domains of HRQL across measures and to compare agreement at different points in time after the stroke event. In addition, we were interested in whether mean proxy-assessed change scores were systematically different from patient scores, an issue important in the analysis of clinical trial data.

Four specific hypotheses were proposed. First, mean cross-sectional HRQL scores assessed by proxy assessment were expected to be lower than patient self-assessment. Second, patient-proxy agreement was expected to be greater at 6 months than at baseline as the patient became more clinically stable. Third, stronger patient-proxy agreement was expected for the more observable domains of HRQL such as mobility, ambulation, and self-care, and poorer agreement was expected on the less observable domains (eg, emotion). Fourth, patient-proxy agreement was expected to be poorest for the visual analog scale (VAS) of the EQ-5D. The EQ-VAS score involves a direct valuation of health. The other summary scores are based on health status assessments and calculated with an algorithm based on community preferences. Thus, VAS scores reflect heterogeneity in both health status and the valuation of health states, whereas the other summary scores reflect only heterogeneity in health states.

**Subjects and Methods**

The study design was an observational longitudinal cohort study that consecutively enrolled hospitalized stroke patients and caregivers who met the selection criteria. Patients were enrolled after the initial acute phase of stroke but before hospital discharge: 95% were enrolled within 2 weeks of stroke. Ischemic stroke was confirmed by CT or MRI brain scan. Cerebral infarctions were classified into 4 subtypes: lacunar infarcts, total anterior circulation infarcts, partial anterior circulation infarcts, and posterior circulation infarcts.

In addition to patient consent, patients were required to have a caregiver (proxy) who also consented to participate. The proxy was a family caregiver such as a spouse or partner, sibling, or offspring or, if unavailable, a friend. Both patient and caregiver had to be able to comprehend English and be ≥18 years of age. Patients were excluded if they had a life expectancy of <6 months for any medical reason, a history of previous degenerative or space-occupying brain disorder, hemorrhagic or lower brainstem stroke, subarachnoid hemorrhage or transient ischemic attack, coma, global or Wernicke’s aphasia, or history of dementia before stroke. Patients and caregivers had to live within 150 km of Edmonton, Alberta, and not be cognitively impaired in the judgment of the clinical assessor. The study was conducted through 2 large teaching hospitals in Edmonton, and ethics approval was obtained from the participating institutions and the Health Research Ethics Review Board at the University of Alberta.

The patient and proxy were requested not to discuss the items with each other during completion of the questionnaires. Assessments were performed at baseline and 1, 3, and 6 months after baseline. After baseline, research assistants contacted and visited the patients and caregivers to oversee questionnaire completion. Research assistants were permitted to assist physically in the completion of the surveys if the respondent requested help.

Standard 1-week recall versions of the HUI questionnaire for self-completion were administered to the patient and proxy and scored as recommended. HUI3 single-attribute utility scores are defined on a scale in which no impairment (ie, normal) is assigned a score of 1.00 and severe impairment (eg, blind on the vision attribute) is assigned a score of 0. Overall scores are on a scale in which perfect health is equal to 1.00 and dead is equal to 0; negative scores imply health states worse than dead. The HUI3 includes vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain.

A standard version of the EQ-5D was administered that comprises a 5-domain health self-classification system and a VAS, described as a “feeling thermometer” rated from 0 to 100, anchored by worst and best imaginable health state. The health state vector from the self-classification system was transformed into a single, index-based preference score (EQ-Index) using the scoring algorithm from York (UK). The EQ-5D was amended for completion from the proxy view of the patient’s health status. Because the HUI questionnaire has standardized instructions for completion by proxy, it preceded the EQ-5D in order of administration.

Differences in central location (mean) of responses by patient and proxy to each dimension of the EQ-5D were tested with a sign test. Agreement on each EQ-5D item was evaluated with $\kappa$, weighted by squared differences. Differences between patient and proxy HUI3 single-attribute scores were assessed with the Wilcoxon rank test (2 tailed). Agreement for HUI3 single-attribute scores and for the overall utility scores was assessed with a 1-way random-effects model–based intraclass correlation coefficient (ICC).

The magnitude of the systematic bias between patient and proxy mean scores was quantified with the standardized response mean (SRM), calculated as patient minus proxy score standardized by the standard deviation of the difference score. Given that the SRM is a variant of effect size ($d$), an absolute standardized difference of $|d| = 0.2$ was interpreted as small effect; $|d| = 0.5$ indicated medium effect; and $|d| = 0.8$ or more was interpreted as large effect.

A generic threshold of discrimination of minimally important differences in HRQL in chronic diseases has been estimated at half an SD. A guideline for interpretation of interrater reliability generalizability coefficients is as follows: poor ($\leq 0.40$), fair ($0.41$ to $0.59$), good ($0.60$ to $0.74$), and excellent ($\geq 0.75$). Results presented focused on baseline and 6-month assessment (further analyses available on request). Because of the nature of the research objectives, item nonresponses were not imputed. Values of $P < 0.05$ were considered statistically significant, and confidence intervals (CIs) were calculated for the 95% level. Statistical analyses were performed with SPSS version 10.1.3 and SAS system version 8.01.

**Results**

Of the 556 patients who presented with potential stroke, 200 patients met the selection criteria and were approached to participate. Of these, 54 patients (27%) declined to participate, and 22 patients (11%) did not participate because of caregiver reluctance (on behalf of the patient or themselves). In addition, 29 patients were ineligible because they did not have an identifiable caregiver (family or friend). The characteristics of the 124 patients and proxy pairs (62% of those eligible) who participated are presented in Table 1. Strokes were image verified by CT or MRI, with initial imaging conducted within 24 hours of admission for >90% of the patients. Two or more scans were performed in 58% of...
At baseline, most patients experienced severe stroke according to Barthel Index scores, with 60% of patients scoring 60 (dependent) and 5% scoring 95 (independent). At 6 months, 15% of the sample had scores 60, and 50% had scores 95.

The number of patient/proxy respondents at each time point was 124/124 (t0), 108/104 (t1), 102/101 (t3), and 98/96 (t6). Of the 26 patients lost to follow-up, 8 patients died. Fewer than 5% of respondents had ≥1 item nonresponse to the EQ-5D and HUI. Differences in dyads available for analysis between the EQ-5D and HUI (Tables 2 through 4) were due to the HUI scoring algorithm, which classifies inconsistent responses for multi-item attributes as missing data.

Proxies demonstrated a central tendency to report more problems than the patient on the EQ-5D at 6 months for self-care, pain/discomfort, and anxiety/depression (P<0.05) (Table 2). Agreement based on κ was good for the more observable dimensions (mobility, self-care) and poor for the less observable dimensions (pain/discomfort, anxiety/depression) at baseline. Exact agreement and ICC point estimates generally improved at the 6-month follow-up, most notably for pain/discomfort.

ICC-based agreement at baseline was fair to good for the more observable attributes of the HUI3 (ambulation, dexterity) and improved at 6 months (Table 3). At both baseline and 6 months, proxies underestimated the extent of problems with hearing compared with patient assessment. Patient self-assessed cognition scores were systematically higher than proxy scores (P<0.05). Poor patient-proxy agreement persisted at 6 months on the attributes of speech, hearing, and cognition. The domain of hearing had a poor ICC-based agreement yet a high level of exact agreement (85%) because

### TABLE 2. Patient and Proxy Responses to EQ-5D Dimensions

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Baseline</th>
<th>6 Months</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Patient (n=124), %</td>
<td>Proxy (n=124), %</td>
</tr>
<tr>
<td><strong>Mobility</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No problems</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>Problems</td>
<td>48</td>
<td>45</td>
</tr>
<tr>
<td>Confined to bed</td>
<td>36</td>
<td>40</td>
</tr>
<tr>
<td><strong>Self-care</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No problems</td>
<td>20</td>
<td>22</td>
</tr>
<tr>
<td>Some problems</td>
<td>41</td>
<td>33</td>
</tr>
<tr>
<td>Unable to wash/dress</td>
<td>38</td>
<td>45</td>
</tr>
<tr>
<td><strong>Usual activities</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No problems</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Some problems</td>
<td>44</td>
<td>36</td>
</tr>
<tr>
<td>Unable to perform</td>
<td>47</td>
<td>56</td>
</tr>
<tr>
<td><strong>Pain/Discomfort</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>44</td>
<td>32</td>
</tr>
<tr>
<td>Moderate</td>
<td>51</td>
<td>63</td>
</tr>
<tr>
<td>Extreme</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td><strong>Anxiety/Depression</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>35</td>
<td>28</td>
</tr>
<tr>
<td>Moderate</td>
<td>61</td>
<td>66</td>
</tr>
<tr>
<td>Extreme</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

*P<0.05; †P<0.01; sign test comparing patient and proxy responses at time point (under proxy heading).
of extreme discrepancy in a small subgroup of patients who reported they were unable hear at all, while their proxies reported the patients had no problems hearing. Dexterity had poor exact agreement but fair ICC-based agreement because, although most patient-proxy responses were not identical, they were generally within 1 response category of each other. Although most patient-proxy responses were not identical, they were generally within 1 response category of each other.

For summary scores, the magnitude of difference between patient and proxy mean scores was absent to small (Table 4). They were generally within 1 response category of each other. Although most patient-proxy responses were not identical, they were generally within 1 response category of each other.

of extreme discrepancy in a small subgroup of patients who reported they were unable hear at all, while their proxies reported the patients had no problems hearing. Dexterity had poor exact agreement but fair ICC-based agreement because, although most patient-proxy responses were not identical, they were generally within 1 response category of each other.

For summary scores, the magnitude of difference between patient and proxy mean scores was absent to small (Table 4). They were generally within 1 response category of each other. Although most patient-proxy responses were not identical, they were generally within 1 response category of each other.

The extent of agreement and systematic differences between patient-proxy assessments of HRQL has practical implications for determining whether proxy assessments can reliably substitute for patient self-assessment in intervention-based studies after stroke. Proxy assessments demonstrated acceptable levels of reliability for cross-sectional estimates of the summary scores of the EQ-Index and HUI3, particularly if obtained ≥1 month after stroke. Proxy assessments of the direct preference-based EQ-5D VAS were less reliable, especially at baseline assessment.

Although only fair agreement was observed between patient and proxy change scores on the EQ-Index and HUI3, the magnitude of systematic differences between assessor types was generally trivial (d<0.20). For instance, mean EQ-5D index-based change scores (t/∆t) were the same for patient assessment (0.32; SD=0.38) and proxy assessment (0.32; SD=0.39). It may be reasonable to contemplate the use of individual proxy assessments to derive change scores for the purpose of group-level inferences against the alternatives: missing data, statistical imputation, or mapping from clinical evaluation.

In general, results were similar to those reported for summary scores in previous studies of the HUI and EQ-5D.4,5 HUI3 single-attribute and overall utility scores displayed a similar pattern of agreement, except that we observed less agreement on the hearing attribute. The fair to good agreement on the EQ-5D observed in the present study at 6 months was comparable to the agreement in a previous study of the

### Table 3. Patient and Proxy HUI3 Single-Attribute Utility Scores

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Baseline*</th>
<th>Baseline†</th>
<th>Patient, Mean (SD)</th>
<th>Proxy, Mean (SD)</th>
<th>Patient-Proxy Difference</th>
<th>Exact Agreement, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean (SD)</td>
<td>Proxy, Mean (SD)</td>
<td>Mean (SD)</td>
<td>ICC (95% CI)</td>
</tr>
<tr>
<td>Vision</td>
<td>0.90 (0.30)</td>
<td>0.85 (0.22)</td>
<td>0.05</td>
<td>0.20</td>
<td>0.43 (0.27–0.53)</td>
<td>67</td>
</tr>
<tr>
<td>Hearing</td>
<td>0.85 (0.31)</td>
<td>0.91 (0.20)</td>
<td>-0.06</td>
<td>0.26</td>
<td>0.47 (0.32–0.60)</td>
<td>81</td>
</tr>
<tr>
<td>Speech</td>
<td>0.83 (0.18)</td>
<td>0.86 (0.17)</td>
<td>-0.03</td>
<td>0.20</td>
<td>0.36 (0.19–0.51)</td>
<td>59</td>
</tr>
<tr>
<td>Ambulation</td>
<td>0.41 (0.36)</td>
<td>0.41 (0.37)</td>
<td>-0.01</td>
<td>0.29</td>
<td>0.69 (0.59–0.78)</td>
<td>43</td>
</tr>
<tr>
<td>Dexterity</td>
<td>0.52 (0.36)</td>
<td>0.59 (0.35)</td>
<td>-0.07</td>
<td>0.33</td>
<td>0.53 (0.40–0.65)</td>
<td>35</td>
</tr>
<tr>
<td>Emotion</td>
<td>0.83 (0.22)</td>
<td>0.82 (0.20)</td>
<td>0.00</td>
<td>0.26</td>
<td>0.26 (0.09–0.42)</td>
<td>48</td>
</tr>
<tr>
<td>Cognition</td>
<td>0.83 (0.29)</td>
<td>0.76 (0.26)</td>
<td>0.07</td>
<td>0.25</td>
<td>0.41 (0.26–0.55)</td>
<td>40</td>
</tr>
<tr>
<td>Pain</td>
<td>0.84 (0.25)</td>
<td>0.69 (0.35)</td>
<td>0.15</td>
<td>0.38</td>
<td>0.14 (0–0.31)</td>
<td>33</td>
</tr>
</tbody>
</table>

*|n=115 to 124.
†|n=91 to 95.
‡|P<0.05; §|P<0.01; §P<0.001, 2-tailed Wilcoxon signed rank test comparing patient and proxy scores.

### Discussion

The extent of agreement and systematic differences between patient-proxy assessments of HRQL has practical implications for determining whether proxy assessments can reliably substitute for patient self-assessment in intervention-based studies after stroke. Proxy assessments demonstrated acceptable levels of reliability for cross-sectional estimates of the summary scores of the EQ-Index and HUI3, particularly if obtained ≥1 month after stroke. Proxy assessments of the direct preference-based EQ-5D VAS were less reliable, especially at baseline assessment.

Although only fair agreement was observed between patient and proxy change scores on the EQ-Index and HUI3, the magnitude of systematic differences between assessor types was generally trivial (d<0.20). For instance, mean EQ-5D index-based change scores (t/∆t) were the same for patient assessment (0.32; SD=0.38) and proxy assessment (0.32; SD=0.39). It may be reasonable to contemplate the use of individual proxy assessments to derive change scores for the purpose of group-level inferences against the alternatives: missing data, statistical imputation, or mapping from clinical evaluation.

In general, results were similar to those reported for summary scores in previous studies of the HUI and EQ-5D.4,5 HUI3 single-attribute and overall utility scores displayed a similar pattern of agreement, except that we observed less agreement on the hearing attribute. The fair to good agreement on the EQ-5D observed in the present study at 6 months was comparable to the agreement in a previous study of the
EQ-5D$^3$ for self-completers who survived at least 3 months after stroke. On all EQ-5D dimensions, proxy assessment was more reliable in stroke compared with dementia patients.$^17$

In considering proxy assessment of the EQ-5D and HUI3 in stroke, the EQ-5D is briefer and simpler for proxies to complete but lacks attributes relevant to stroke that are included in the HUI3 such as cognition, speech, and dexterity. Interestingly, the attribute of cognition demonstrated poor to moderate agreement between patient and proxy assessment, with proxy scores being systematically lower at both baseline and 6 months. This discrepancy points to the need to contemplate the validity of a different perspective such as the family caregiver as an additional criterion for evaluating the usefulness of proxy respondents for neurologically compromised conditions. Disagreement is not necessarily undesirable, and multiple viewpoints can be valid and informative.$^{18}$ Caregivers may recognize functional limitations that patients are unaware of or deny. Arguably, assessments of HRQL that are used to inform decision making in health care based on community-based preferences need not be restricted to the patient perspective. Generalizability was attenuated by the exclusion of patients who lacked identifiable informal caregivers and thus likely to have less social support. The study sample was composed of fewer patients with mild stroke compared with the distribution of stroke types described in larger, comprehensive studies of stroke.$^9$ The tertiary care hospitals in the study are referred the more serious stroke cases in the region, and all participants were hospitalized for at least 1 day. The sample was well suited for studying the research question, but study results cannot be generalized to cognitively impaired stroke patients, for whom proxy assessments are especially salient. Finally, the HUI3 scoring algorithm potentially introduced a bias that favored more agreement because respondents with illogical response sets were filtered out.

In conclusion, we found that patient-proxy agreement using the EQ-5D and HUI3 was comparable to previous studies even though the sample consisted of relatively fewer patients with mild stroke. Results suggested that proxies are more reliable for assessing stroke patients using community preference-based summary scores of generic HRQL measures (eg, HUI3, EQ-5D Index) and that proxy assessments of direct preference-based scores (EQ-VAS) were less reliable. Patient-proxy assessments had greater agreement if performed ≥1 month after the stroke event. Sequential proxy assessments to obtain change scores is not recommended but may be considered against the alternatives such as statistical imputation or mapping from clinical evaluation.

**Acknowledgments**

We are grateful to AstraZeneca for an unrestricted grant. The funding agency played no role in the design, interpretation, or analyses of the project and did not review or approve the manuscript. At the time of the study, Dr Pickard was supported by studentships from the Health Research Foundation/Canadian Institutes for Health Research (HRF/CIHR) and Alberta Heritage Foundation for Medical Research (AHFMR). Dr Johnson holds a Population Health Investigator award with AHFMR and a Canada Research Chair in Diabetes Health Outcomes. Dr Feeny holds a CIHR-Rx&D Chair. We acknowledge the contributions of Alison Weingardt, Aldis Hunt, and Kendra Jones. We also acknowledge the helpful comments of Dr Denis Revicki. Note that David Feeny has a proprietary interest in...
Health Utilities Incorporated, the firm that distributes copyrighted Health Utilities Index materials.

References
Agreement Between Patient and Proxy Assessments of Health-Related Quality of Life After Stroke Using the EQ-5D and Health Utilities Index
A. Simon Pickard, Jeffrey A. Johnson, David H. Feeny, Ashfaq Shuaib, K.C. Carriere and Abdul M. Nasser

*Stroke*. 2004;35:607-612; originally published online January 15, 2004; doi: 10.1161/01.STR.0000110984.91157.BD

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://stroke.ahajournals.org/content/35/2/607

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in *Stroke* can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to *Stroke* is online at:
http://stroke.ahajournals.org/subscriptions/