ELSEVIER

Patient Education and Counseling 50 (2003) 151–155

Patient Education and Counseling

www.elsevier.com/locate/pateducou

Evaluating the reliability and validity of three tools to assess the quality of health information on the Internet

Gbogboade Ademiluyi[a], Charlotte E. Rees[b,*], Charlotte E. Sheard[a]

[a]*Division of Psychiatry, University of Nottingham, Nottingham, UK*
[b]*Institute of Clinical Education, Peninsula Medical School, Tamar Science Park, ITTC Building, Davy Road, Plymouth PL6 8BX, UK*

Received 18 February 2002; received in revised form 12 May 2002; accepted 10 June 2002

Abstract

The quality of Internet information needs to be evaluated and several tools exist for this purpose. However, none have demonstrated reliability and validity. This study tested the internal consistency and validity of the information quality tool (IQT), quality scale (QS) and DISCERN using 89 web sites discussing smoking cessation. The inter-rater reliability of the tools was established by exploring the agreement between two independent raters for 22 (25%) of the sites. The IQT and DISCERN possessed satisfactory internal consistency (as measured by Cronbach's α). The IQT, QS and DISCERN showed satisfactory inter-rater reliability (as measured by κ and intraclass correlations). The IQT, QS and DISCERN correlated positively with each other, supporting the convergent validity of the tools. This study provides some evidence for the reliability and validity of the IQT, QS and DISCERN, although this needs testing in further research with different types of Internet information and larger sample sizes.

Keywords: Reliability; Validity; Assessment tools; Internet information

1. Introduction

Health information has many benefits for patients and their families. It can increase individuals' knowledge of their disease and its treatments [1], can aid coping [2], reduce distress and anxiety [3], help individuals make informed decisions regarding their treatment [4] and can increase individuals' adherence with medical advice, e.g. screening regimes [5]. Increased access to the Internet has provided patients with a new source of information and the rapid growth of the Internet has triggered an information revolution [6]. An example of this revolution is the explosion of cancer support groups (CSGs) on the Internet, which have increased dramatically over recent years [7].

Several authors have attempted to define what constitutes quality health information on the Internet. Kim et al. [8] conducted a review of published criteria for evaluating health-related web sites and identified 29 published journal articles and rating tools (e.g. HON code) that had explicit criteria for assessing health-related web sites. They extracted 165 criteria from the tools and articles, 132 (80%) of which were grouped under 12 specific categories and 33 (20%) were grouped as miscellaneous because they were unique or lacked specificity. The most frequently cited quality criteria were those dealing with content, design and aesthetics of the site, disclosure of authors, sponsors or developers, currency of information, authority of source, ease of use and accessibility and availability.

Using these quality criteria, several researchers [9–11] have shown that health information on the Internet is of variable quality. Impicciatore et al. [11] evaluated the reliability of Internet information on managing fever in children. The authors conducted a systematic search of web pages relating to the home management of feverish children using two search engines; Yahoo and Excite. They then evaluated this information using a self-prepared checklist including the type of organisation that created the web site; the country it operated from and the language in which the information was offered. The authors also considered more specific items relating to fever and its management, e.g. the minimum temperature considered as fever. The information provided on the web sites was compared with evidence-based guidelines to parents on managing fever at home [12]. The authors retrieved 41 web pages but found that only a few sites

* Corresponding author. Tel.: +44-1752-764447; fax: +44-1752-764226.
E-mail address: charlotte.rees@pms.ac.uk (C.E. Rees).

0738-3991/02/$ – see front matter
PII: S0738-3991(02)00124-6

provided a complete and accurate picture of managing the feverish child at home. Impicciatore et al. [11] concluded that there was an urgent need to check patient information on the Internet for completeness, accuracy and consistency. Silberg et al. [9] also concluded that medical information on the Internet contains too much incomplete, misleading and inaccurate information, which could result in harmful effects for both patients and healthcare professionals who fail to use the Internet properly [6].

Although many authors agree on key criteria for assessing the quality of Internet information, few tools possess any type of rating scale. The exceptions to this include instruments like the information quality tool (IQT) [13], quality scale (QS) [14] and DISCERN [15], which have all been documented in recent literature. However, none of these measures have undergone rigorous testing to determine their reliability or validity with Internet information. This study attempts to address this gap in the research literature by evaluating the reliability and validity of these three tools.

2. Methods

2.1. Site selection

As part of a study to evaluate the quality of smoking cessation information on the Internet, we conducted 40 separate searches for smoking cessation information in October and November 2001. We used four search terms (''quitting smoking'', ''stopping smoking'', ''giving up smoking'', and ''smoking cessation'') with each of 10 popular search engines (AltaVista, AOL Search, AskJeeves, Excite, Google, Hotbot, Looksmart, Lycos, MSN Search, and Yahoo). With the exception of the four Yahoo searches (which yielded less than 10 web sites), the top-10 listed web sites were identified for each of the searches, leading to an initial sample of 370 web sites. Of this initial sample, 281(75.9%) sites were discarded, either because they were duplicated sites, dead links, did not contain smoking cessation information, or they required a fee for accessing the information. This resulted in a sample of 89 unique web sites for analysis in this study.

2.2. Assessment tools

2.2.1. Information quality tool (IQT)

GA used the 21-item IQT [13] to evaluate the quality of smoking cessation information on the Internet. This scale includes items relating to authorship (items 1–7), sponsorship (items 8–10), currency (items 11–13, 16), accuracy (items 14–15, 17), confidentiality (item 18), and navigability (items 19–21). Although each item requires a yes or no answer, items are weighted according to their importance. Items perceived to be most important are given a weight of 1 (e.g. ''Are the site author's credentials listed?'') and the three items weighted 1 must be answered 'yes' for the site to pass. Those perceived to be least important are given a weight of 0.036 (e.g. ''Is a search engine provided?''). The total score for the scale can range from 0 to 4. CES independently evaluated the quality of 22 (25%) of the 89 web sites using the IQT to determine its inter-rater reliability.

2.2.2. Quality scale (QS)

GA used the seven-item quality scale [14] to evaluate the quality of smoking cessation information. This scale includes items relating to ownership (item 1), authorship (item 2), source (item 3), currency (item 4), interactivity (item 5), navigability (item 6), and balance (item 7). Each item is accompanied by a three-point Likert scale where 0 is failure to satisfy the criteria for that item, 1 means to partially satisfy, and 2 to completely satisfy the criteria for that item. The total score for the scale can range from 0 to 14. CER independently evaluated the quality of 22 (25%) of the 89 web sites using the quality scale to determine its inter-rater reliability.

2.2.3. DISCERN

GA used the 16-item DISCERN tool [15] to evaluate the quality of Internet information on treatment choices for smoking cessation. The first section (questions 1–8) evaluates the reliability of the information (e.g. ''Is it clear what sources of information were used to compile the publication?'') and the second section (questions 9–15) considers the quality of the information on treatment choices (e.g. ''Does it describe the benefits of each treatment?''). Five-point Likert scales ranging from 1 (no) to 5 (yes) accompany these items. The final section (question 16) assesses the overall rating of the publication on a five-point Likert scale ranging from 1 (low quality with serious or extensive shortcomings) to 5 (high quality with minimal shortcomings). We also calculated a total score by summing the scores for items 1–15. This gave a score ranging from 15 to 75, with low scores indicating poor quality and high scores indicating good quality. CER independently evaluated the quality of 22 (25%) of the 89 web sites using the DISCERN tool to determine its inter-rater reliability.

2.3. Statistical analyses

Data were analysed using SPSS (version 10) or SAS (release 6.12). The internal consistency of the scales ($n = 89$) was established using Cronbach's alpha (α) coefficients. The index of agreement between the two raters for each item of the scales ($n = 22$) was established by kappa (κ) coefficients (IQT) or weighted κ coefficients (QS and DISCERN). The index of agreement between the two raters for the scales' total scores was determined by intraclass correlation coefficients. The convergent validity of the scales ($n = 89$) was assessed by establishing the relationship between the scales' total scores using Spearman's rho correlation coefficients.

Table 1
Summary of agreement between raters for each item of the IQT

IQT item	κ	P value	Level of agreement[a]
1	0.812	<0.001	Almost perfect
2	0.831	<0.001	Almost perfect
3	0.353	0.030	Fair
4	0.741	0.001	Substantial
5	0.582	0.006	Moderate
6	0.353	0.030	Fair
7	0.673	0.001	Substantial
8	0.137	0.203	Slight
9	0.450	0.035	Moderate
10	1.000	<0.001	Perfect
11	0.817	<0.001	Substantial
12	0.463	0.010	Moderate
13	_[b]	_[b]	_[b]
14	0.680	0.001	Substantial
15	0.899	<0.001	Almost perfect
16	0.273	0.127	Fair
17	0.553	0.007	Moderate
18	0.482	0.024	Moderate
19	0.197	0.228	Slight
20	1.000	<0.001	Perfect
21	1.000	<0.001	Perfect

[a] Level of agreement as indicated by Landis and Koch [16].
[b] Could not be calculated because three cells were empty but the agreement was 100%.

3. Results

3.1. Reliability

The internal consistency of the IQT was Cronbach's $\alpha = 0.634$. The inter-rater agreement for the IQT items ranged from $\kappa = 0.137$ (item 8) to 1.000 (items 10, 20 and 21) (Table 1). The level of agreement between the total IQT scores, as measured by an intraclass correlation coefficient was 0.543 ($P = 0.004$).

The internal consistency of the QS was Cronbach's $\alpha = 0.413$. The inter-rater agreement for the QS items ranged from $\kappa = 0.209$ (item 5) to 0.708 (item 4) (Table 2). The level of agreement between the total QS scores, as measured by an intraclass correlation coefficient was 0.759 ($P < 0.001$).

The internal consistency of DISCERN was Cronbach's $\alpha = 0.777$. The inter-rater agreement for the DISCERN items ranged from $\kappa = 0.102$ (item 10) to 0.761 (item 5) (Table 3). The index of agreement for the overall quality rating (item 16) was $\kappa = 0.516$. The level of agreement between the total DISCERN scores, as measured by an intraclass correlation coefficient was 0.823 ($P < 0.001$).

Table 2
Summary of agreement between raters for each item of the QS

QS item	Weighted κ	CI (95%)	Level of agreement[a]
1	0.234	−0.114, 0.582	Fair
2	0.443	0.074, 0.812	Moderate
3	0.593	0.304, 0.881	Moderate
4	0.708	0.460, 0.955	Substantial
5	0.209	−0.147, 0.566	Fair
6	0.267	−0.024, 0.558	Fair
7	0.532	0.286, 0.778	Moderate

[a] Level of agreement as indicated by Landis and Koch [16].

Table 3
Summary of agreement between raters for each item of DISCERN

DISCERN item	Weighted κ	CI (95%)	Level of agreement[a]
1	0.460	0.243, 0.678	Moderate
2	_[b]	_[b]	_[b]
3	0.484	0.214, 0.754	Moderate
4	0.705	0.451, 0.958	Substantial
5	0.761	0.561, 0.961	Substantial
6	0.655	0.446, 0.864	Substantial
7	0.434	0.188, 0.679	Moderate
8	0.549	0.306, 0.793	Moderate
9	0.541	0.331, 0.750	Moderate
10	0.102	−0.099, 0.303	Slight
11	0.507	0.217, 0.798	Moderate
12	0.211	−0.008, 0.429	Fair
13	_[b]	_[b]	_[b]
14	0.380	0.078, 0.681	Fair
15	_[b]	_[b]	_[b]
16	0.516	0.249, 0.784	Moderate

[a] Level of agreement as indicated by Landis and Koch [16].
[b] Could not be calculated because three cells were empty but the agreement was 20% (item 2), 95% (item 13) and 68% (item 15).

3.2. Convergent validity

With the exception of the correlation between the IQT total score and the DISCERN overall quality rating, each tool correlated positively and significantly with every other tool (Table 4).

4. Discussion

The internal consistency of the IQT was found to be moderate in this study (Cronbach's $\alpha = 0.634$). This Cronbach's α coefficient did not meet the requirements for

Table 4
Relationships (Spearman's rho correlation coefficients) between total scores for each scale and the DISCERN overall quality rating

	QS total score	DISCERN total score	DISCERN overall quality rating
IQT total score	$r = 0.518$, $P < 0.001$	$r = 0.243$, $P = 0.039$	$r = 0.226$, $P = 0.057$
QS total score	*	$r = 0.529$, $P < 0.001$	$r = 0.378$, $P = 0.001$
DISCERN total score	*	*	$r = 0.801$, $P < 0.001$

satisfactory reliability as stated by Bland and Altman [17] (i.e. >0.70). This may be due to the fact that all of the items had dichotomous response categories rather than Likert scales [18]. Although κ coefficients ranged from 0.137 to 1.000, the majority of calculable kappas ($n = 14/20$, 70%) had coefficients between 0.41 and 1.00, indicating moderate, substantial or perfect levels of agreement [16]. Generally, high levels of agreement were found for more objective items (e.g. 'Is a search engine provided?'—item 20) and lower levels of agreement for more subjective items (e.g. 'Is the site easily navigable and presented in an organised manner?'—item 19). The intraclass correlation coefficient was found to be moderate in this study (0.543, $P = 0.004$), providing tentative evidence for the reliability of the IQT.

The internal consistency of the QS was poor in this study (Cronbach's $\alpha = 0.413$), a value well below the benchmark deemed acceptable by Bland and Altman [17]. A possible explanation for this low α is that the QS contains only seven items with three-point Likert scales [18]. Although, weighted κ coefficients ranged from 0.209 to 0.708, the majority of items ($n = 4/7$, 57.1%) had coefficients between 0.41 and 0.80, indicating moderate or substantial levels of agreement [16]. Again, higher levels of agreement were found for more objective items (e.g. currency—item 4) and lower levels of agreement for more subjective items (e.g. navigability—item 6). The intraclass correlation coefficient was found to be satisfactory in this study (0.759, $P < 0.001$), providing some evidence for the reliability of the QS.

The internal consistency of DISCERN was satisfactory in this study (Cronbach's $\alpha = 0.777$) [17]. Although, weighted κ coefficients ranged from 0.102 to 0.761, the majority of calculable kappas ($n = 10/13$, 76.9%) had coefficients between 0.41 and 0.80, indicating moderate or substantial levels of agreement [16]. Furthermore, the index of agreement for the overall quality rating (item 16) was moderate ($\kappa = 0.516$). Again, higher levels of agreement were found generally for more objective items (e.g. 'Is it clear what sources of information were used to compile the publication?'—item 4), supporting the findings of earlier research [19]. Furthermore, lower levels of agreement were found mainly for more subjective items (e.g. 'Is it relevant?'—item 3). However, there were a few exceptions to this rule with some objective items regarding treatment options yielding lower κ values (e.g. items 10, 12 and 14). These low levels of agreement probably resulted from the raters interpreting the item hints (or instructions) differently. This was easily done as the item instructions were written for treatments for diseases rather than treatments for lifestyle changes such as smoking cessation. Nevertheless, the weighted κ coefficients were similar to those reported by Charnock et al. [15]. The intraclass correlation coefficient was found to be satisfactory in this study (0.823, $P < 0.001$), providing evidence for the reliability of DISCERN.

With the exception of the correlation between the DISCERN overall quality rating and the IQT total score ($r = 0.226$, $P = 0.057$), each tool correlated positively and significantly with every other tool. This suggests that the tools were measuring similar concepts (i.e. the quality of Internet information) and therefore supports the convergent validity of the tools.

However, a number of methodological limitations must be taken into consideration when interpreting these results. Although we conducted a comprehensive search with 10 popular search engines and four different search terms, this yielded a sample of only 89 unique web sites. In addition, due to time and financial constraints, only 22 web sites were used to determine the inter-rater reliability of the quality tools. These sample sizes both fall short of the 100 cases Kline [20] suggested was necessary to minimise the standard error of reliability statistics. Furthermore, this sample size may not have been large enough to detect significant relationships between the scales' total scores, leading to Type II errors. This may have been the case for the correlation between the IQT total score and the DISCERN overall quality rating, which demonstrated a trend towards a statistically significant relationship ($P = 0.057$).

Furthermore, because we only used smoking cessation web sites to test the reliability and validity of these tools, it is difficult to know whether these tools will be reliable and valid with Internet information discussing other health-related issues. Therefore, we urge researchers to test the reliability and validity of these tools with Internet information discussing other types of health information.

Finally, these tools assess aspects of quality that are valued by healthcare professionals, which may be different from those aspects valued by some patients. For example, information posted by patients and patient support groups may be viewed as good quality by patients because it is reassuring and useful, but the same information might be rated as poor quality by healthcare professionals because it is inaccurate or misleading. It is important that future research examines the broader and more patient-centred concept of quality as well as healthcare professionals' quality criteria.

4.1. Practice implications

This study provides some evidence for the reliability and validity of the IQT, QS and DISCERN, although this needs testing in further research evaluating different types of Internet information and with larger sample sizes. It is important that healthcare professionals inform patients of the variable quality of health information on the Internet. Healthcare professionals could use tools like the IQT, QS and DISCERN to evaluate the quality of health information on the Internet, so that they are better able to direct patients to higher quality web sites. Furthermore, healthcare professionals could educate patients about the existence of these tools. Indeed, tools like DISCERN were designed for use by healthcare professionals and patients, so doctors could encourage their patients who regularly access Internet information to use them.

References

[1] Humphris GM, Duncalf M, Holt D, Field EA. The experimental evaluation of an oral cancer information leaflet. Oral Oncol 1999;35:575–82.

[2] Harrison-Woermke DE, Graydon JE. Perceived informational needs of breast cancer patients receiving radiation therapy after excisional biopsy and axillary node dissection. Cancer Nurs 1993;16:449–55.

[3] Michie S, Rosebert C, Heaversedge J, Madden S, Parbhoo S. The effects of different kinds of information on women attending an outpatient breast clinic. Psychol Health Med 1996;1:285–96.

[4] Davison BJ, Kirk P, Degner LF, Hassard TH. Information and patient participation in screening for prostate cancer. Patient Educ Couns 1999;37:255–63.

[5] Myers RE, Chodak GW, Wolf TA, Burgh DY, McGrory GT, Marcus SM, et al. Adherence by African American men to prostate cancer education and early detection. Cancer 1999;86:88–104.

[6] Jadad A, Gagliardi A. Rating health information on the Internet: navigating to knowledge or to Babel? J Am Med Assoc 1998; 279: 611–4.

[7] Klemm P, Hurst M, Dearholt SL, Trone SR. Cyber solace: gender differences on Internet cancer support groups. Comput Nurs 1999;17:65–72.

[8] Kim P, Eng TR, Deering MJ, Maxfield A. Published criteria for evaluating health-related web sites: review. Br Med J 1999;318: 647–9.

[9] Silberg WM, Lundberg GD, Musacchio RA. Assessing, controlling, and assuring the quality of medical information on the Internet. J Am Med Assoc 1997;277:1244–5.

[10] Winker MA, Flanagin A, Chi-Lum B, White J, Andrews K, Kennett RL, et al. Guidelines for medical and health information sites on the Internet: principles governing AMA web sites. J Am Med Assoc 2000;283:1600–6.

[11] Impicciatore P, Pandolfini C, Casella N, Bonati M. Reliability of health information for the public on the world wide web: systematic survey of advice on managing fever in children at home. Br Med J 1997;314:1875–9.

[12] EI-Radhi AS, Carroll J. Management of fever. In: EI-Radhi AS, Carroll J, editors. Fever in paediatric practice. Oxford: Blackwell Scientific Publications, 1994. p. 229–31.

[13] Mitretek Systems. Information quality tool. http://hitiweb.mitretek.org/iq/questions.asp (accessed 7 May 2002).

[14] Sandvik H. Health information and interaction on the Internet: a survey of female urinary incontinence. Br Med J 1999;319:29–32http://www.bmj.com (accessed 7 May 2002).

[15] Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999;53:105–11http://www.discern.org.uk (accessed 7 May 2002).

[16] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:139–74.

[17] Bland JM, Altman DG. Cronbach's alpha. Br Med J 1997;314:571.

[18] Leon AC, Marzuk PM, Portera L. More reliable outcome measures can reduce sample size requirements. Arch Gen Psychiatry 1997;52:867–71.

[19] Rees CE, Ford JE, Sheard CE. Evaluating the reliability of DISCERN: a tool for assessing the quality of written patient information on treatment choices. Patient Educ Couns 2002;47: 273–5.

[20] Kline P. The handbook of psychological testing. London: Routledge, 2000.