Chapter 8
An Information Reliability Index as a Simple Consumer-Oriented Indication of Quality of Medical Web Sites

Federico Cabitza

Abstract. Since typical healthcare consumers may lack sufficient knowledge to evaluate the reliability of health-related contents published online, recent researches are addressing the usefulness of Web page evaluation tools to help these consumers assess the quality of the indications they retrieve online. This paper contributes in this line by proposing an intentionally simple composite index of information quality, the so called Medical Information Reliability (MIR) index. This index takes the attitudes of potential and actual consumers toward information quality into account, and it is intended to be applied to online sources of medical information as "trust indicator" to provide their potential consumers with a simple percentage score by which to evaluate the reliability of what they are consulting. The main idea underlying this index is to consider information quality a multidimensional aspect of an online resource and relate it to the extent such a resource is compliant with explicit requirements formulated by third-party endorsement bodies. The method to calculate the MIR index on a sample of medical sites is presented in a step-by-step manner, and a user study is discussed that validated its application to the domain of the Alternative and Complementary Medicine.

1 Background and Motivations

Users rely on online resources in regard to their health for a number of reasons: e.g., to see if others complain their same symptoms and see how these had their disorders solved (especially in case of sensitive or socially stigmatized illnesses); to find the actual meaning of unfamiliar terms that had been used by healthcare

Federico Cabitza
Università degli Studi di Milano-Bicocca, Viale Sarca 336, 20126 Milano, Italy
e-mail: cabitza@disco.unimib.it

G. Pasi et al. (Eds.): *Qual. Issues in the Management of Web Information*, ISRL 50, pp. 159–177.
DOI: 10.1007/978-3-642-37688-7_8

professionals in previous encounters; to sift opinions on the effectiveness of alternative treatments or on the reliability of a healthcare provider; to find support groups or just someone to exchange ideas and experiences with about their own health [1]. Looking for healthcare-related advice and information on the Web is easy, fast and extremely cheap, especially in comparison to getting access to the often well remunerated opinion of a doctor; this is the reason why this is a common behavior: approximately two thirds of adult population both in the US and in Europe claim to use the Internet for health care purposes [2]; this is also why an increasing number of people claim to ground their healthcare decisions on what they find on the Internet [3]. Our work lies in the research field aimed at providing final consumers and end-users of Web sites with a simple way to assess the reliability of the information they get access to online. In particular, we propose a methodology by which the content published by an online resource can be rated according to a series of standard domain-specific criteria, and be associated with a simple percentage indicator, the Medical Information Reliability (MIR) index; this index is aimed at making consumers aware of the extent online content has been produced according to quality-related guidelines. The main motivation backing this proposal, first presented in [4], lies in the need to address quality assessment from the consumer perspective and to make quality indicators intentionally simple to understand for lay consumers in critical domain like healthcare is.

Thus, our first motivation lies on the fact that for a consumer of health-related information to be able to assess the Information Quality (IQ) of an indication found on the Internet is particularly important [5]; this is especially true in light of two aspects: first, recently a number of research studies that reviewed Web sites providing healthcare-related information have detected that these sites presented several inaccuracies in their content; this finding raises serious concerns about the IQ that health consumers can encounter on the Internet [6–8]. Second: typical consumers of healthcare-related information online have only a limited knowledge of what they are reading (otherwise it is likely that they would not be seeking medical information through that means) and therefore they could be unable to judge its quality and reliability with full confidence by their own [9].

The second motivation for our proposal lies in the fact that the characteristics of individual consumers and their past interactions with a content provider in general are not sufficient to assess the actual reliability of a health-related Web site [10]; therefore there is a need for mechanisms of trust verification that are based on social institutions and intermediaries. To this aim, a number of initiatives have been conceived to help health information consumers seek, find and access high-quality information: these initiatives include gateway sites (portals), evaluation instruments and codes of conduct associated with some "surface markers" [11]. In particular, Chang and Cheung [12] have showed that third-party certifications are the most effective way for a Web site to gain trust from its prospective consumers when its reputation is unknown. The simplest mechanisms of this type are the so called "surface markers" and "trust indicators"; in particular, the mechanism that is based on a "code of conduct" that is developed by a third party organization and that is associated to a label or logo is called "kite marking" [13]; a kite mark is like a "seal of quality" that a specific Web site can display if it is declaring to abide by the terms of such a code and if it is periodically found compliant with the code guidelines. By means of this kind of trust indicator, visitors of Web sites can get an idea of the IQ of an online resource [14], in the assumption that a certified trustworthy provider, thanks to its internal policies of IQ control, would always publish reliable contents [11]; in this case, then, an indication of trustworthiness is used as a predictor of accuracy [9,15]. Due to their immediacy and easiness to use, a growing number of organizations have recently developed codes of conduct that are associated to an evaluation and certification service that assigns kite marks in the healthcare domain (e.g., the HON code).

In its simplest terms, our proposal is a method to attach a percentage-based "kite mark" to a Web site, the so called MIR index; this is made according to i) a subjective evaluation of the compliance of the Web site to a set of domain specific codes of conduct (by a trained set of raters), and to ii) a consumer-oriented prioritization of the aspects that these codes of conduit regulate on a more or less prescriptive level. In this paper we will present next the original contribution, discussed in light of the relevant related work; then we will present a stepwise method to calculate the MIR index for a specific web resource; finally we will present a user study we undertook to validate its application to a specific healthcare domain, the domain of the Alternative and Complementary Medicine; to this aim, we will present the method adopted, the results and finally, we will discuss them in light of the main objectives outlined in the next section.

2 The Medical Reliability Index Score

The Medical Reliability Index is a *weighted composite index* that we conceived as an evaluation tool and systematic method whose output is a numerical "trust indicator", to provide the users of an online health-related resource with a simple indication of its "level of reliability" and, hence, of the degree of IQ of the content published therein. In the proposal of the MIR index we have been driven by three main requirements, which we drew from the specialist literature regarding the kind of third-party certification mentioned in the previous section.

1. *Focus on the patients' (information) needs.* According to [16], the service quality of electronic resources can not prescind from research initiatives that focus on the relative importance assigned by service consumers to different quality dimensions and perceived attributes. To this regard, it has been shown that consumers of health-related information usually develop a personal perception of how accurate they believe a content is by relying on visual elements, like layout, color schemes and icons, which are displayed by the online resource, rather than on content [17]. Moreover, as noted in [18] and [19] in regard to the IQ of the content available online, it is important to distinguish

between the needs of patients and the needs/expectations of healthcare professionals [20]: although for both categories of content consumers there is a need for a rigorous assessment of the quality of health-related websites [21], patients' needs are reported to value "trustworthiness" more than "availability" and "accessibility" [22], which are the main concerns of doctors. As the profile of the typical consumer of health-related content is changing over time, with patients and laymen becoming the main consumers of this kind of offer, it is recognized the increasing need to concentrate more on patients [23]. This calls for visual indicators of IQ that are "simple" and "straightforward" explicitly, since it is also reported that the majority of health information seekers do not check IQ-related indications in a consistent manner, like date of publication and original source of the information [24].

2. *Compliancy to the so called "codes of conduct"*. These codes are proposed by a number of endorsement bodies and associations with the aim to guarantee the generic public of health-related Web sites that get the corresponding certification of compliance; these codes may differ with respect to several aspects, like the intended target population, the specific scope, and the declared and actual aims [25]. A recent survey has found that no code of conduct can be considered universally suitable to evaluate the IQ of different health websites [21]. This calls for the requirement that a good indicator must be a *composite* one, that is one that takes existing complementary IQ indicators and aggregates them together in some consistent and systematic way.
3. *Keep it simple, but not simpler.* Although trust indicators and quality seals, e.g., kite marks, are simple to understand even for laymen and straightforward in their meaning (simplistically put, if the indicator is present, then the "site is OK"; otherwise, nothing can be said about its IQ), it is noticed that they may give a false sense of security [26]. This calls for the requirement of avoiding dichotomic measures (e.g., good vs. bad quality mark, pass vs. no pass certification), but rather to adopt a numerical and percentage-oriented approach that could provide consumers with a more precise, yet still qualitative, indication of the extent the resource is compliant with domain-specific guidelines for IQ assurance.

In light of these three requirements, we devised the MIR score as a composite, percentage-like index whose numerical expression is obtained through a systematic eight-step evaluation method to express the extent IQ-related criteria have been met by a specific health-related information provider (online resource, or Web site in what follows) with respect to the ideality, i.e., 100% of the identified criteria being satisfied. More technically speaking, the MIR index is calculated according to Formula 1, where for each of the *n* IQ dimensions taken in consideration, w_i is the ranked weight for the i-th dimension d_i (e.g., accuracy); and c_k is the k-th IQ criterion associated to d_i (i.e., such that $F(c_k) = d_i$ [1]), and evaluated for a specific Web site, *s*; tc_i is the total number of criteria to be met in regard to dimension d_i. See Table 1 for more details on the meaning of the variables involved in Formula 1.

[1] Or, also, the pair $(c_k, d_i) \in$ M, that is a set introduced in Table 1.

$$MIR(s) = \frac{\sum_{i=1}^{n}\left(w_i \bullet \left(\frac{\sum_{k=1}^{m} c_k(s)}{tc_i}\right)\right)}{\sum_{i=1}^{n} w_i} \qquad (1)$$

The fact of presenting the MIR score as a percentage is obviously aimed at facilitating laymen consumers in understanding "how reliable a content provider is" with respect to a conventional "upper limit" (i.e., all requirements met), which in a traditional "kite mark" approach would be associated to the issuing of a single trust indicator. Beside indicating also a partial compliance of a given online resource with respect to the available best practices (i.e., absolute benchmarking), the fact that the MIR score is numeric allows also for paired-sample comparisons and, consequently, for its adoption as an (internal) audit tool for the continuous improvement of the IQ of health-related content: i.e., on the one hand, it enables the homogeneous comparison of different online resources (i.e., relative benchmarking) and hence their ranking in online directories, gateway portals or search engines; on the other hand, it enables the progressive evaluation of a single resource over time, and hence to detect trends in IQ policies and actual performance.

3 The Evaluation Process Related to the MIR Index

Besides simplicity, another element that is worthy of note in regard to MIR as an evaluation tool it is its capability to be "tailored" to different needs, aims and domains, as the case study that we will present on the Complementary and Alternative Medicine will show. This aspect derives from the recognition, mentioned above and reported also in [21], that a truly universal and semi-automatic evaluation tool would be overambitious and probably practically infeasible. Thus, Formula 1 presents two variables, namely the weights by which IQ dimensions are prioritized (i.e., *w*) and the criteria by which IQ is assessed (i.e., *c*) that can either be set once and for all; or be object of tailorization according to the actual uses that are intended for the MIR score (e.g., site valorization, benchmarking, trend analysis, continuous IQ improvement, information retrieval). In this light, IQ dimensions (i.e., a third parameter, *d*, that does not show up in Formula 1) are just a user-centered way to prioritize IQ success criteria, that is a way to take the users' perceptions and preferences into account [16] (cf. the first requirement mentioned above). This called for the conceptual separation between the evaluation tool (and related score) and the evaluation process; in its turn, the latter one can be further distinguished into a phase of "adaptation" (or inspection [32]) of the tool, where criteria by which a health-related site is considered reliable (or not) and their weights are uniquely identified and set; and a phase of "use" of the tool, where Web sites are manually checked against the above identified criteria and a numerical score is attached to these sites at a given time.

This process can be further articulated in a stepwise manner by identifying eight distinct tasks:

1. Identification of the IQ Dimensions involved
2. Identification of the IQ Criteria involved
3. Criteria Categorization
4. Prioritization of the IQ Dimensions
5. Weight Definition
6. Site review
7. Score calculation
8. Score dissemination

In Table 1, we describe this evaluation process in some details and indicate, for each step listed above, some techniques that can be adopted for its execution, and the intended outputs.

Table 1 The evaluation process toward the definition of a MIR score for a generic online resource

Step No.	Description	Technique(s) involved	Step Outputs
1	IQ Dimension Identification, and definition / characterization of each IQ dimension in simple but unambiguous terms.	User study (survey); Focus group (Delphi method); Literature review; or a combination of these.	A set D, of *n* IQ dimensions: D = {d_1, …, d_n}. E.g.: d_1 is Completeness, d_2 is Accuracy, d_3 is Timeliness.
2	IQ Criteria Identification and characterization to the original formulation expressed by third-party endorsement providers.	Literature review; Focus group (Delphi method); or a combination of both.	A set C of *m* success criteria: C = {c_1, …, c_m}, with c_i being a Boolean function that evaluates the i-th criterion, i.e., a single requirement by which to assess the quality of a resource, such as: c_i : S → {0, 1}, with S set of web sites (s) under review. E.g., given a web site $s_{i,}$ c_1: "in s_i , is the source of information always identified?"; $c_{2:}$ "in s_i , is the contact information for the site administrator displayed?"; c_3: "in $s_{i,}$ are medical and other disclaimers posted and easily accessible?".

Table 1 (*continued*)

3	Categorization of the criteria defined in Step 2 in terms of the dimensions defined at step 1.	Categorization through inspection; Coding through reliable keyword matching (cf. content analysis and inter-coder reliability assessment).	A collection M, of all ordered pairs M: $\{(c_i, F(c_i))\}$ with i from 1 to m with F: $C \rightarrow D$, i.e., a function by which a single success criterion is associated with a single IQ dimension. In other words, we obtain a list of dimensions operationally defined in terms of IQ specifications. E.g., see Table 2.
4	IQ dimension prioritization	User study; Focus groups (Delphi method) or a combination of these.	A total order $\succ \subseteq$ D X D (cf. D in step 1) by which $d_1 \succ d_2 \succ \ldots d_{n-1} \succ d_n$; in other words, we obtain an ordinal ranking of the IQ dimensions found in step 1. E.g.: 1) Accuracy; 2) Completeness; 3) Timeliness.
5	Weight definition and assignment to ordinal ranks.	Literature review; Introspection; Focus groups (Delphi method), or a combination of these.	An ordered set W of weights, $W = \{w_1, \ldots, w_n\}$, where w_i is to be associated with a specific d_k, with i = k.
6	Review of an online resource (s) and check of its content against the criteria defined at Step 2.	Evaluation by a (pool of) trained expert(s), be it either extensive or upon a random sample of pages from a web site (s).	A collection E, of all ordered pairs E: $(s, C_i(s))$ with C_i defined at step 2. In other words, by applying all the c_i in C to s, we obtain a list of dichotomic evaluation scores (1/0), one for each IQ criterion.
7	MIR score calculation	See Formula 1	The MIR score for s at time T_i .
8	MIR score dissemination	Site Directory (aka gateway providers); Kite Mark; or both.	

In light of the process outlined in Table 1, two more points are worthy of note. First, although seemingly redundant, it is important to distinguish between Step 6, i.e., the criterion-by-criterion review of an online medical resource, and Step 7, i.e., the to some extent “mere” calculation of the MIR score that follows this review. This is because the evaluation of a Web site with respect to its compliance with the identified IQ criteria is conceptually, as well as operationally, a different task from associating such a review with a numerical score. This latter task could be repeated for different sets W of weights in order to choose the optimal one that, e.g., makes important differences between, e.g., competing sites more manifest; obviously in this case, there would be no need to replicate the review; or Step 6 could be assigned to a pool of evaluators that, in a similar way to the coding task of Step 3, are collaboratively called to reach a consensus on what criteria are really met in all those cases this is not a trivial task but rather something that requires experience and interpretative skills[2].

Second: although strictly stepwise and linear in its overall structure, the MIR evaluation process is intrinsically iterative in all of the steps of the adaptation phase, from Step 1 to Step 5 (see Figure 1). All these steps can encompass a collaborative process in which, respectively relevant IQ dimensions, success criteria and weights are defined in a progressive manner through increasing levels of consensus within a group of prospective users, analysts or domain experts. Moreover, the overall process is intended to loop from Step 8 back to Step 6 (the Use phase in Figure 1) to enable continuous IQ improvement and benchmarking, once the IQ dimensions, criteria and weights set in the adaptation phase have been held constant, of course.

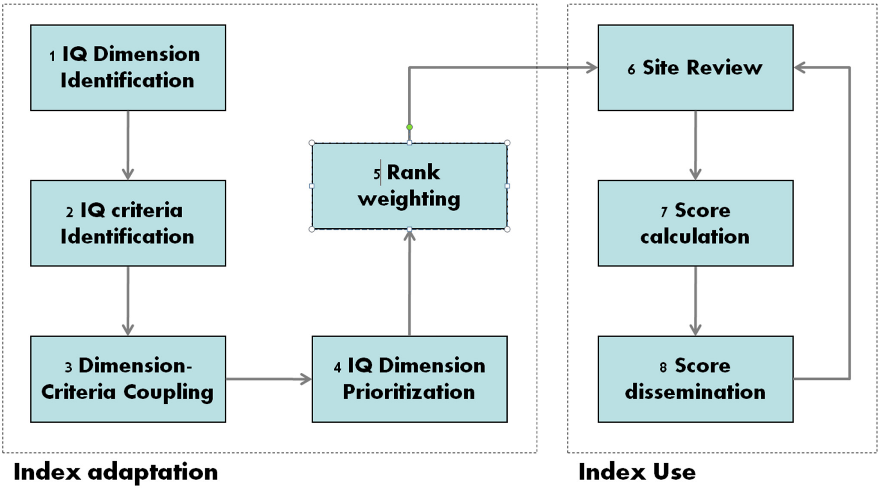
1 IQ Dimension Identification
2 IQ criteria Identification
3 Dimension-Criteria Coupling
4 IQ Dimension Prioritization
5 Rank weighting
6 Site Review
7 Score calculation
8 Score dissemination
Index adaptation
Index Use

Fig. 1 Graphical representation of the process of the MIR index adaptation and use

4 Validation of the MIR Index in the Medical Domain

In the Introduction we have already made the point of how, due to the intrinsic heterogeneity and extent of the World Wide Web, the quality of health information that a consumer could find online can differ a lot [27]; and how this can be

[2] In any case, step 7 can be executed only if set W and E (see table 1) are available.

related to legitimate concerns that trusting low quality (or simply not verified) content on health-related matters could have even serious consequences on the consumers' health [28]. This is specially true in those ambits of medicine where there is still a lack of institutional roles acting for consumers as reliable providers of information, like doctors and pharmacists; this is the case of Non Conventional, Complementary and Alternative Medicines, that is the field that is usually referred with the acronym NCM/CAM (in what follows just CAM for simplicity's sake). For this reason, we decided to deploy the MIR evaluation tool in this specific domain and proceed with a preliminary validation that could address the evaluation of the reliability of some Web sites that provide their customers with advices, indications, results from the specialist literature and market news about CAM-related products and remedies in the Italian and English speaking contexts.

The National Center for Complementary and Alternative Medicine (NCCAM) defines CAM as the broad set of healing remedies and resources that are either complementary or alternative to those established within the conventional healthcare practice in a particular society [29]. According to the World Health Organization, more than half of the European citizens have used CAM remedies and the expenditure trend on this kind of medicines, therapies and practices is increasing year by year [30].

4.1 Validation Method

To validate the adaptability of the MIR index to the CAM domain and its feasible application to related Web sites, we designed a user study that could follow the step-wise process outlined in Table 1. This study used a mixed methodology with multiple data sources. In particular, in the first step we analyzed the pertinent literature to identify the main dimensions along which IQ in CAM Web resources is usually measured; the same approach was followed in Step 2 to identify a set of criteria that should be met by a Web site to be considered a reliable source of health-related information. Step 3 was conducted by a small panel of coders following the tenets of content analysis [31] and software engineering inspection methods [32]. The prioritization task involved in Step 4 was conducted in virtue of the results of an exploratory empirical user study; in this study, we surveyed a convenience sample of 101 healthcare information consumers on their attitudes and quality expectations for CAM-related information also on the basis of their previous experience with the pursuit and retrieval of such information on the Internet [33]. The participants have been invited to respond to an online questionnaire where they were given the set of IQ dimensions identified in Step 1 and were asked to assess the perceived importance of each dimension with respect to their tasks of seeking and consulting CAM-related information on an ordinal scale ("very important", "important", "moderately Important"; "of little importance") [9]. On the basis of these subjective assessments, we produced a ranking of the dimensions in order of *perceived importance*. Finally, we have tested the applicability of this metrics by applying the MIR index to a convenience selection of Web sites that publish CAM-related content and hence drew a first indication of their reliability according to our proposal.

4.2 Results

A comprehensive literature survey (including, e.g., [8,10,14,34–39]) allowed us to realize that there is no single list of IQ dimensions or attributes that could be considered the "gold reference" for healthcare website evaluation. For instance, Kim et al. reviewed 29 rating tools presenting explicit criteria to assess health-related Web sites: the most frequently cited criteria regard currency of information, authority of source, ease of use, accessibility/availability, disclosure of authors and content accuracy [37]. Eysenbach, in a systematic review on how health website are evaluated in the specialist literature, found that the most frequently used quality criteria regard accuracy, completeness, readability, design, disclosures, and references provided [8]. More recently, Stvilia et al. [14] analysed thousands of e-mail communication instances in the IPL's Q&A service archives from 2005 to 2007 and identified seven IQ criteria to be relevant to healthcare IQ judgments: accuracy, authority, completeness, currency, objectivity, relevancy, and understandability. In 2009, O'Grady [38] developed an evaluation framework for health Web sites that encompasses the IQ dimensions of: content accuracy, credibility, completeness, understandability, relevance, level of personalization, privacy, security, usability, and accessibility; even more recently (2011), Tao et al. [20], on the basis of a user study focusing on the perspective of healthcare consumers, identified a taxonomy of IQ attributes encompassing understandability, completeness, reputation, adequacy of reference, relevancy, accuracy, site reputation among others. In light of these and other studies, we defined a list of six IQ dimensions that could cover the main quality-related aspects with CAM information published online; for each of these dimensions, we formulated a definition with no ambition of sound comprehensiveness, but rather with the aim to help the coders involved in Step 3 reach a sufficient level of agreement, as well as to provide the participants of the empirical study accomplished in Step 4 with a common ground and shared definition of the terms used in the questionnaire.

The IQ dimensions we adopted for this study are:

(i) *accuracy*, expressed as 'the extent a piece of information is true and reliable according to either reality or a "gold standard" reference' (e.g., a medical dictionary or textbook, a scientific paper);

(ii) *completeness*, defined as 'the extent a piece of information is reported in a complete way to inform its consumers according to their needs';

(iii) *accessibility*, as 'the extent a piece of information is easy to be found and consulted' (cf. availability);

(iv) *currency*, as 'the extent a piece of information is up-to-date';

(v) *usefulness*, as 'the extent a piece of information is easy to understand (cf. understandability) and apply to a specific context or need';

(vi) *authority*, as 'the extent the author or source of a given piece of information is known and considered trustworthy'.

For Step 2, a second literature survey was aimed at identifying the main criteria CAM-related Web site must meet to obtain a third-party certification of the quality of its content (e.g.,[40] [10, 11]). We identified three codes of conduct that could fit our aims well:

1) The so called "HONcode", which is issued by the Health On the Net Foundation [12] in terms of a list of eight general requirements that a Web site must satisfy on a yearly basis to get and maintain the related certification.
2) The Web Feet Health Criteria for Site Selection[3], a detailed collection of criteria that was part of a larger collection of indications to retrieve high quality information on the Internet for school, business and library purposes.
3) The checklist issued by the NCCAM, which encompassed ten "questions" addressing as many aspects to consider to judge a Web site a reliable source of CAM-related information (or not).

These evaluation tools were chosen mainly on the basis of the recent and comprehensive literature review reported in [11]: we included the HONcode and the Webfeet collection as these ones were found to cover a superset of the IQ-related aspects addressed by most of the existing other evaluation instruments; and also because they were found to be the most different and hence complementary ones, in terms of rank correlation. We adopted also the NCCAM checklist as this was found to be the only one specifically designed for the CAM domain. In doing so, we were confident to extract from these tools all the main recurring aspects that are covered by the evaluation instruments available online [11], while, at the same time, to also consider the most specific instrument for the domain at hand, and therefore take into account all the relevant aspects or success criteria related to IQ mentioned in the literature. The resulting 42 criteria extracted in Step 2 of our validation study are reported in Table 2.

Table 2 The list of criteria selected from the literature survey. The list is used to evaluate c_k in Formula 1. A requirement is intended to be either satisfied for a site s ($c_k(s) = 1$) or not satisfied ($c_k(s) = 0$), according to an evaluation/judgment task performed by a trained rater.

Evaluation Tool	Criteria Checklist
NCCAM	1. Is who runs this site explicitly reported or otherwise clear? 2. Is Who pays for the site explicitly reported or otherwise clear? 3. Is the purpose of the site explicitly reported or otherwise clear? 4. Is Where the information comes from explicitly reported or otherwise clear? 5. Is What the basis is of the information explicitly reported or otherwise clear? 6. Is How the information is selected explicitly reported or otherwise clear? 7. Is How current the information is explicitly reported or otherwise clear? 8. Is How the site chooses links to other sites explicitly reported or otherwise clear? 9. Is What information about you the site collects (and why) explicitly reported or otherwise clear? 10. Is How the site manages interactions with visitors explicitly reported or otherwise clear?

[3] This tool, now apparently discontinued, can be found at the following URL: `http://www.webcitation.org/5QFjclQjk`

Table 2 (*continued*)

HONcode	1. Authoritative. Are the qualifications of the authors clearly stated? 2. Complementarity: Is it clear that information provided should support, not replace, the doctor-patient relationship? 3. Privacy: Is the privacy and confidentiality of personal data submitted to the site by the visitor fully respected? 4. Attribution: Are the source(s) of published information, date medical and health pages properly cited? 5. Justifiability: Are claims relating to benefits and performance properly backed up? 6. Transparency: Is content presented in an accessible way, and contact information accurate? 7. Financial disclosure: Are all funding sources properly identified and acknowledged? 8. Advertising policy: Is advertising content clearly distinguished from editorial content?
Web Feet Health	1. Is it true that Source of information is identified? 2. Is it true that The contact information for the source or site administrator is displayed? 3. Is it true that The expertise and reputation of the source are considered? 4. Is it true that The expertise and reputation of the site's host are considered? 5. Is it true that The information is not easily available at other sources? 6. Is it true that Reviewers (clinicians, subject-area experts, and researchers) make every effort to ensure that the information is free of errors? 7. Is it true that The information and images are objective, balanced, and unbiased? 8. Is it true that The information has sufficient scope to cover the topic for the intended audience? 9. Is it true that The information is readable and free of spelling and grammatical errors? 10. Is it true that Sponsorship is clearly indicated, and advertising is minimal? 11. Is it true that Medical and other disclaimers are posted? 12. Is it true that Site is updated frequently, typically indicated by a recent "last updated" date? 13. Is it true that Pages list the date of the most recent update and/or the dating of the information is made clear in an accessible area of the site? 14. Is it true that Links work, and they are relevant and appropriate? 15. Is it true that The site loads in a reasonably short time? 16. Is it true that The site is easy to access and navigate? 17. Is it true that Navigation includes clear headings and intuitive icons, menus, and directional symbols that foster independent use? 18. Is it true that Standard multimedia formats such as HTML are used? 19. Is it true that Most information is accessible without special plug-ins such as Adobe Acrobat Reader? 20. Is it true that Logical options are available for printing and downloading all or selected text or graphics? 21. Is it true that The site follows good graphic design principles? 22. Is it true that Information for specific audiences, such as consumer information within a professional site, is easy to locate? 23. Is it true that The site has a text size that is easy to read for the intended audience? 24. Is it true that Product advertising is not intrusive?

In regard to Step 3, we enrolled three coders (including the author) and provided them with the list of the IQ dimensions (and related definitions) detected at Step 1 as the shared "codebook" by which they were called to classify the IQ criteria independently of each other [41]. A score of inter-rater reliability was then calculated by means of the KALPHA macro by Andrew F. Hayes for SPSS v. 17.0, obtaining a Krippendorff's Alpha score of 0.69. This value is usually associated with a less than optimal reliability, and therefore with exploratory conclusions only (as it is below the conventional threshold for high reliability, i.e., K ≥ 0.8 [42]); nevertheless, this result made us confident that a representative coding scheme could be eventually found by the coders involved in a subsequent phase, when the resulting pairs criteria-dimension were openly discussed in a Delphi-like manner [43]. In Table 3, we report the result of this collaborative task of characterization of the IQ criteria found in Step 2 in terms of the IQ dimensions identified in Step 1.

Table 3 The definition of each IQ dimension regards what was agreed upon by coders in step 3 as well as the definition given to participants in step 4

IQ Dimension	Definition	Coding of criteria
Accuracy (tc_d: 6 criteria)	This dimension mainly relates to the requirement that the site should not contain commissions, i.e., misleading statements likely to cause physical harm [13].	NCCAM: 5 HONcode: 4 Web Feet Health: 3, 6, 7, 9
Completeness (tc_d: 11 criteria)	This dimension mainly relates to the requirement that the site should not contain omissions, i.e., vital information that should have been mentioned and that All claims should be justified with appropriate references to scientific sources.	NCCAM: 2, 4, 5, 6 HONcode: 5, 7 Web Feet Health: 1, 2, 8, 10, 13
Accessibility (tc_d: 16 criteria)	This dimension mainly relates to how easy it is to find a content that is pertinent to one's own needs, as well as to retrieve it again over time; thus, this dimension also relates to the persistence of the content itself and to its uniquely identifiability.	NCCAM: 5, 10 HONcode: 8, 4, 5, 6 Web Feet Health: 2, 14, 15, 16, 17, 18, 19, 21, 22, 23
Currency (tc_d: 2 criteria)	This dimension mainly relates to how timely a new content is produced after that a related scientific evidence has been produced and published in the specialist literature, that is the extent the content site is up-to-date with respect to the available knowledge.	NCCAM: 7 HONcode: None Web Feet Health: 12
Usefulness (tc_d: 8 criteria)	This dimension mainly relates to the extent a site presents content that is understandable, interesting, and therefore valuable for the intended consumers; this dimension also relates to the extent an intended consumer can take advantage of the information consulted, that is how easily a piece of information is applicable to either everyday needs or more specific information requirements.	NCCAM: 3 HONcode: 2, 8 Web Feet Health: 8, 11, 16, 20, 22

Table 3 (*continued*)

Authority (tc_d: 7 criteria)	This dimension mainly relates to the extent the provider is considered reliable, trustworthy and able to satisfy information needs of its customers, and consequently, how easy it is to trace to the author or provider and assess its reliability.	NCCAM: 1, 4 HONcode: 1, 4 Web Feet Health: 1, 4, 5

To perform Step 4, we conducted an online survey that involved a convenience sample of healthcare information consumers that were questioned about their attitudes and quality expectations for CAM-related information.

The survey was conceived as a Computer Assisted Web Interview (CAWI) that was delivered through an online questionnaire platform (Limesurvey[4]). Participants were recruited among acquaintances and colleagues of the author and students of his classes, and were invited to join the study either through a personal email or being forwarded to the survey page through posts published on the main social networks (i.e., Facebook, MySpace e Twitter); word of mouth did the rest. The questionnaire was kept open for 18 days and closed on September 2011, when 101 completed forms had been collected. In Table 4 we report the demographic data extracted from the sample of respondents and the segmentations that were performed for the inference statistical study.

Table 4 Selected demographic information of the respondent sample involved in the empirical study

Characteristic	Options	Responses %
ICT skills	Elementary or basic	39.6
	Advanced or expert	60.4
Interest toward CAM	Very low or low	39.6
	High or very high	60.4
Knowledge about CAM	Very Poor or poor	69.3
	Good or very good	30.7
Frequency for CAM	From Never to Sometimes	68.4
	Frequently or Very frequently	31.7

In Table 5 we report the results coming from the user study. These regard, for each IQ dimension, the average ranking and the ordinal category that better represents the average attitude toward that dimension (i.e., the median of the response distribution). The average rank for each IQ dimension has been calculated counting how many times that dimension was considered the most important among the other ones, how many times the second one, the third one and so on; and then calculating the arithmetic mean of the total score. This is just one of the ways in which large samples of respondents can be challenged about relative rankings without asking them for a ranking directly, so as to minimize acquiescence bias; other techniques can be obviously adopted, as one from those reviewed in [44].

[4] `http://www.limesurvey.org/`

Table 5 Average Ranking from the empirical study for each IQ dimension

IQ Dimension	Overall Rank[5]	Median Perception[6]
Accuracy	1.56	Very important
Completeness	1.60	Very important
Accessibility	2.07	Very important
Currency	2.22	Very important
Usefulness	2.43	Important
Authority	2.96	Important

Due to the convenience-driven recruitment, the user study employed in Step 4 presents the common limitation of not being based on a sample that is fully representative of the target population, i.e., potential consumers of CAM-related information. To limit non response bias, we stratified the sample into subgroups by age, education, familiarity with ICT (ICT skills in Table 4), knowledge and interest toward CAM, frequency with which information on CAM remedies is either sought or consulted (see Table 4). We associated the subgroups with dichotomic variables (the options column in Table 4) and performed a Mann-Whitney U test (since the assessments were performed on an ordinal scale) for each specific IQ dimension: no significant difference at a 95% confidence level was found among these groups with regard to their assessments of the perceived importance of the IQ dimensions. This fact, as well as the relatively large number of respondents, does not eliminate the bias due to accidental sampling, but makes the results consistent with the requirements of marketing research [45], i.e., suitable to detect attitudes and preferences in potential consumers of CAM-related information.

According to the ranking derived from the user study accomplished in Step 4, we adopted one of the simplest weighting function and assigned each IQ dimension to its corresponding weight, starting from Accuracy ($w_d = 6$, in Formula 1) to Authority ($w_d = 1$), with unitary decrements.

Subsequently, we proceeded in Step 6: in this step, the list of criteria reported in Table 2 is intended to be consulted by a neutral rater[7] or by anyone specifically instructed to check whether the *n*-th criterion, associated with the *k*-th IQ dimension, is actually met by the Web site under evaluation, or not. We then reviewed five Web sites that we selected on a convenience basis among those that were publishing CAM-related information on a daily basis at the time of the validation and we calculated the MIR score for each of these online resources. In Table 6 we report the results from this purposely exemplificatory evaluation.

[5] Smaller number indicates higher importance. Values are close since we did not forced the respondents to chose a rank for each IQ dimension explicitly (to minimize random bias) but we derived this indication from their single assessments.

[6] The median values of the distribution of "perceived importance" that are reported in the rightmost column are equal to the modes for that variable, i.e., the value that has been chosen by the majority of the respondents.

[7] In the case at hand, the author made the review.

Table 6 MIR scores applied reviewing five CAM-related web sites, visited in Summer 2010

Web Site	MIR index score
Italiasalute	42.0%
Viveremeglio	42.0%
Mentalhelp	75.9%
Wiki4cam	51.6%
Altmeds	62.2%

5 Discussion and Concluding Remarks

In this paper we have presented the Medical Information Reliability (MIR) index, a composite weighted score of IQ intended to facilitate healthcare consumers in the task of judging the reliability of an online resource in lack of sufficient knowledge to perform this task without external visual aids.

With respect to other post-hoc IQ evaluation instruments that have been proposed with similar purposes in the healthcare domain [11], the MIR index is novel for its modularity, simplicity and consumer-centredness. First, the MIR index can integrate multiple IQ criteria from instruments that are issued and maintained by various certification bodies. This integration requires only to associate each new dichotomous criterion with the pertinent IQ dimension. Also the ranking weights can be adjusted over time to better fit either local or specific target readerships. Second, the MIR index is purposely conceived as a simple percentage indication of the extent a Web site is compliant with the best practices and guidelines for IQ assurance, where 100% indicates a fully compliant site. As such, it is a tool for health-related information consumers to support them in getting an idea of the reliability of a source of content published in the Web; it is also a tool for gateway providers [11] and, potentially, search engines to refer visitors to better online resources and benchmark them; and it is also a tool for Web sites managers, maintainers and owners, as a means to achieve and guarantee continuous improvement in the eyes of their customers. Lastly, the MIR index is innovative for the idea to include the concept of "requirement prioritization" in a synthetic score: this concept, which is borrowed from the requirement engineering field [32], has inspired the ranking of homogeneous groups of criteria in terms of more understandable meta-level concepts, i.e., the concept of IQ dimension, and suggested to base such a ranking on the actual attitude of potential consumers of health-related information. As part of the further research that is needed to assess the actual value of such a tool, we applied the evaluation process and resulting score to a panel of Web sites publishing consumer-oriented content periodically in the field of the Complementary and Alternative Medicine. This domain was chosen not only because it is receiving strong interest by an increasing population of consumers, but also because the lack of institutional roles and bodies (at least in Europe) that could issue certified indications and proven evidences of effectiveness from the field makes the development and testing of evaluation tools that could contribute in improving the reliability of online resources a pressing need and an interesting challenge in the agenda of both Academic and professional research.

Acknowledgements. The author would like to gratefully acknowledge the work of Marco Buonvino, who contributed during his Master Degree Thesis in both the definition of the MIR score and in accomplishing the most part of the user study.

References

1. Sillence, E., et al.: How do patients evaluate and make use of online health information? Social Science & Medicine 64(9), 1853–1862 (2007)
2. Atkinson, N.L., et al.: Using the Internet for Health-Related Activities: Findings From a National Probability Sample. Journal of Medical Internet Research 11(1), e4 (2009)
3. Baker, L.: Use of the Internet and E-mail for Health Care Information: Results From a National Survey. JAMA: The Journal of the American Medical Association 289(18), 2400–2406 (2003)
4. Cabitza, F.: Introducing a Composite Index of Information Quality for Medical Web Sites. In: Quality of Life thorugh Quality of Information - Proceedings of the 24th Conference of the European Federation for Medical Informatics, MIE 2012, August 26-29, Pisa, Italy (2012) (forthcoming)
5. Gustafson, D.: Evaluation of ehealth systems and services. BMJ 328(7449), 1150–1150 (2004)
6. Bernstam, E.V., et al.: Commonly cited website quality criteria are not effective at identifying inaccurate online information about breast cancer. Cancer 112(6), 1206–1213 (2008)
7. Silberg, W.M., et al.: Assessing, Controlling, and Assuring the Quality of Medical Information on the Internet: Caveant Lector et Viewor–Let the Reader and Viewer Beware. JAMA 277(15), 1244–1245 (1997)
8. Eysenbach, G.: Empirical Studies Assessing the Quality of Health Information for Consumers on the World Wide Web: A Systematic Review. JAMA 287(20), 2691–2700 (2002)
9. Xu, Y., et al.: Relevance judgment: What do information users consider beyond topicality? Journal of the American Society for Information Science and Technology 57(7), 961–973 (2006)
10. Bailey, B.P., et al.: An examination of trust production in computer-mediated exchange. In: Proceedings of the 7th Conference on Human Factors and the Web (2001)
11. Breckons, M., et al.: What Do Evaluation Instruments Tell Us About the Quality of Complementary Medicine Information on the Internet? JMIR 10(1), e3 (2008)
12. Chang, M.K., et al.: Online Trust Production: Interactions among Trust Building Mechanisms [Internet], p. 181c. IEEE (cited January 27, 2012)
13. Delamothe, T.: Quality of websites: kite marking the west wind. British Medical Journal 7, 843–844 (2000)
14. Stvilia, B., et al.: A model for online consumer health information quality. Journal of the American Society for Information Science and Technology 60(9), 1781–1791 (2009)

15. Spink, A., et al.: From highly relevant to not relevant: examining different regions of relevance. Information Processing & Management 34(5), 599–621 (1998)
16. Zeithaml, V.A., et al.: Service quality delivery through web sites: A critical review of extant knowledge. Journal of the Academic of Marketing Science 30, 362–375 (2002)
17. Stanford, J., et al.: Experts vs. online consumers: A comparative credibility study of health and finance web sites. Consumer Reports WebWatch (2002)
18. Potts, H.W.W., Wyatt, J.C.: Survey of Doctors' Experience of Patients Using the Internet. Journal of Medical Internet Research 4(1), e5 (2002)
19. Sillence, E., et al.: Trust and mistrust of online health sites. In: Proceedings of CHI 2004, pp. 663–670 (2004)
20. Tao, D., et al.: Consumer Perspectives on Quality Attributes in Evaluating Health Websites [Internet]. IEEE (2012)
21. Hanif, F., et al.: The role of quality tools in assessing reliability of the Internet for health information. Informatics for Health and Social Care 34(4), 231–243 (2009)
22. Pletneva, N., et al.: Results of the 10th HON survey on health and medical Internet use [Internet]. Health on the Net Foundation, Geneva (2010)
23. Bliemel, M., Hassanein, K.: Consumer satisfaction with online health information retrieval: A model and empirical study. E-Service Journal 5(2), 53–84 (2007)
24. Fox, S.: Online Health Search 2006 [Internet]. PEW Internet & American Life Project (2006)
25. Baur, C., Deering, M.J.: Proposed frameworks to improve the quality of health websites: review. Medscape General Medicine 26(e35) (2000)
26. Hanif, F., et al.: The quality of information about kidney transplantation on the World Wide Web. Clinical Transplantation 21(3), 371–376 (2007)
27. Kalichman, S.C.: Quality of Health Information on the Internet. JAMA: The Journal of the American Medical Association 286(17), 2092–2095 (2001)
28. Schmidt, K., Ernst, E.: Assessing websites on complementary and alternative medicine for cancer. Annals of Oncology 15(5), 733–742 (2004)
29. Goldstein, M.S.: The growing acceptance of complementary and alternative medicine. In: Bird, C.E., Conrad, P., Fremont, A.M. (eds.) Handbook of Medical Sociology, pp. 284–297. Pearson Education Limited, Harlow (1999)
30. Roberti di Sarsina, P., Iseppato, I.: Looking for a Person-Centered Medicine: Non Conventional Medicine in the Conventional European and Italian Setting. Evidence-Based Complementary and Alternative Medicine, 1–8 (2011)
31. Krippendorff, K.: Content analysis: An introduction to its methodology. Sage, Thousand Oaks (2004)
32. Holzinger, A.: Usability Engineering for Software Developers. Communications of the ACM 48(1), 71–74 (2005)
33. Holzinger, A., et al.: The effect of previous exposure to technology on acceptance and its importance in usability and accessibility engineering. Universal Access in the Information Society 10(3), 245–260 (2010)
34. Eysenbach, G., Koehler, C.: How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. BMJ 324(7337), 573–577 (2002)
35. Anderson, K.A., et al.: A Systematic Evaluation of Online Resources for Dementia Caregivers. Journal of Consumer Health on the Internet 13, 1–13 (2009)
36. Lorence, D., Abraham, J.: A study of undue pain and surfing: using hierarchical criteria to assess website quality. Health Informatics Journal 14, 155–173 (2008)

37. Kim, T.R., Deearing, M.J., Maxfield, A.: Published criteria for evaluating health related web sites: Review. British Medical Journal 318, 647–649 (1999)
38. O'Grady, L., et al.: Measuring the Impact of a Moving Target: Towards a Dynamic Framework for Evaluating Collaborative Adaptive Interactive Technologies. Journal of Medical Internet Research 11, 9 (2009)
39. Charnock, D., et al.: DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. Journal of Epidemiology & Community Health 53(2), 105–111 (1999)
40. Goldschmidt, P.G.: A Report on the Evaluation of Criteria Sets for Assessing Health Web Sites [Internet]. Health Improvement Institute and Consumer Reports WebWatch (2003)
41. Carletta, J.: Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics 22, 249–254 (1996)
42. Krippendorff, K.: Reliability in Content Analysis. Some Common Misconceptions and Recommendations. Human Communication Research (3), 411–433 (2004)
43. Okoli, C., Pawlowski, S.D.: The Delphi method as a research tool: an example, design considerations and applications. Information & Management 42(1), 15–29 (2004)
44. Valadares Tavares, L.: A model to support the search for consensus with conflicting rankings: Multitrident. International Transactions in Operational Research 11(1), 107–115 (2004)
45. Fricker, R.D., Schonlau, M.: Advantages and Disadvantages of Internet Research Surveys: Evidence from the Literature. Field Methods 14(4), 347–367 (2002)