Info without Side Effects (iWISE): Development of a validated lay checklist for assessing online health information - a mixed-methods study

Ursula Griebler, Christina Kien, Irma Klerings, Benedikt Lutz, Eva Krczal, Dominic Ledinger, Iris Mair, Robert Emprechtinger, Filiz Keser Aschenberger, Bernd Kerschner

Submitted to: Journal of Medical Internet Research
on: December 05, 2024

Table of Contents

Original Manuscript 5
Supplementary Files 38
Figures 39
Figure 1 40
Multimedia Appendixes 41
Multimedia Appendix 1 42

Info without Side Effects (iWISE): Development of a validated lay checklist for assessing online health information - a mixed-methods study

Ursula Griebler[1]; Christina Kien[1]; Irma Klerings[1]; Benedikt Lutz[2]; Eva Krczal[3]; Dominic Ledinger[1]; Iris Mair[1]; Robert Emprechtinger[4]; Filiz Keser Aschenberger[5]; Bernd Kerschner[1]

[1]Department for Evidence-based Medicine and Evaluation University for Continuing Education Krems Krems an der Donau AT
[2]Department for Knowledge and Communication Management University for Continuing Education Krems Krems an der Donau AT
[3]Department for Economy and Health University for Continuing Education Krems Krems an der Donau AT
[4]QUEST Center for Responsible Research Berlin Institute of Health at Charité (BIH) Berlin DE
[5]Department for Continuing Education Research and Educational Technologies University for Continuing Education Krems Krems an der Donau AT

Corresponding Author:
Ursula Griebler
Department for Evidence-based Medicine and Evaluation
University for Continuing Education Krems
Dr. Karl-Dorrek-Str. 30
Krems an der Donau
AT

Abstract

Background: The internet has become a major source of health information, yet the quality of online health information varies considerably. Users' ability to evaluate the trustworthiness of online health information is limited, as around half of Europeans have limited health literacy. Existing checklists and tools are either prepared for research purpose use or use by health care professionals, have not been developed by considering the lay user perspective, are too long and complicated to be used by laypersons, or were developed for written health information.

Objective: To develop and validate a checklist that enables laypersons to evaluate the trustworthiness of online health information without prior training.

Methods: We employed a multistage mixed-methods approach including: (1) a comprehensive literature review to identify existing tools and quality criteria; (2) an expert Delphi study with six specialists in patient communication and health information; (3) cognitive interviews with 19 lay users in two rounds; (4) application testing on 15 selected health information webpages with 20 additional lay users; (5) a determination of the factual correctness of 100 health information webpages by assessing the difference between the claimed and factual strength of evidence on a health information webpage; and (6) validation testing by research team members on 100 health information webpages using a Bayesian logistic regression model to analyze the predictive validity. In the final step, we integrated all quantitative and qualitative results to select final checklist items.

Results: From an initial pool of 1,740 items extracted from 73 documents, we systematically condensed the list through multiple evaluation and testing rounds. To ensure the checklist is user-friendly, we involved a diverse group of potential users. The final product, named the Info without Side Effects (iWISE) checklist, contains seven items that assess key aspects of health information trustworthiness, including the absence of advertising, balanced presentation of information, limited use of professional jargon, origination from an independent organization, citation of sources, mention of scientific validation, and presence of a publication date. The checklist demonstrated the ability to distinguish between evidence-based and non-evidence-based online health information during validation testing with a nearly 100% probability that the health information is correct if all items were marked with yes.

Conclusions: The iWISE checklist represents a user-friendly, validated tool for evaluating online health information trustworthiness. With only seven items, it is easy to remember and could significantly improve critical health literacy. Future research should test its reliability for social media posts and health information videos.

(JMIR Preprints 05/12/2024:69529)

DOI: https://doi.org/10.2196/preprints.69529

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Info without Side Effects (iWISE): Development of a validated lay checklist for assessing online health information - a mixed-methods study

Abstract

Background: The internet has become a major source of health information, yet the quality of online health information varies considerably. Users' ability to evaluate the trustworthiness of online health information is limited, as around half of Europeans have limited health literacy. Existing checklists and tools are either prepared for research purpose use or use by health care professionals, have not been developed by considering the lay user perspective, are too long and complicated to be used by laypersons, or were developed for written health information.

Objective: To develop and validate a checklist that enables laypersons to evaluate the trustworthiness of online health information without prior training.

Methods: We employed a multistage mixed-methods approach including: (1) a comprehensive literature review to identify existing tools and quality criteria; (2) an expert Delphi study with six specialists in patient communication and health information; (3) cognitive interviews with 19 lay users in two rounds; (4) application testing on 15 selected health information webpages with 20 additional lay users; (5) a determination of the factual correctness of 100 health information webpages by assessing the difference between the claimed and factual strength of evidence on a health information webpage; and (6) validation testing by research team members on 100 health information webpages using a Bayesian logistic regression model to analyze the predictive validity. In the final step, we integrated all quantitative and qualitative results to select final checklist items.

Results: From an initial pool of 1,740 items extracted from 73 documents, we systematically condensed the list through multiple evaluation and testing rounds. To ensure the checklist is user-friendly, we involved a diverse group of potential users. The final product, named the Info without Side Effects (iWISE) checklist, contains seven items that assess key aspects of health information trustworthiness, including the absence of advertising, balanced presentation of information, limited use of professional jargon, origination from an independent organization, citation of sources, mention of scientific validation, and presence of a publication date. The checklist demonstrated the ability to distinguish between evidence-based and non-evidence-based online health information during validation testing with a nearly 100% probability that the health information is correct if all items were marked with yes.

Conclusions: The iWISE checklist represents a user-friendly, validated tool for evaluating online health information trustworthiness. With only seven items, it is easy to remember and could significantly improve critical health literacy. Future research should test its reliability for social media posts and health information videos.

Introduction

With its accessibility and vast resources, the internet has become an increasingly popular source of health information for the general public. Individuals frequently turn to online platforms to seek answers to their health-related questions, research symptoms, and explore treatment options. A recent study in the European Union reported that 56% of individuals aged 16 to 74 used the internet for health-related information-seeking [1], with the highest rates up to 83% in Finland, 78% in the Netherlands, and 75% in Norway. In Austria, 76% used the internet in 2020 to search for information on health issues [2].

While the abundance of health information available online can be empowering, it also presents significant challenges. The quality and reliability of this information vary widely, ranging from evidence-based medical advice to potentially harmful misinformation [3]. This disparity raises concerns about lay users' ability to discern trustworthy sources from unreliable ones. The latest European health literacy survey showed that more than half of users have difficulties in deciding whether the information they found online is trustworthy and objective [4]. According to a survey by the Bertelsmann Foundation [5], between 64% and 84% of participants in Germany rely on advertisement-funded or scientifically unverified web content. Available evidence-based websites were only known to 18% to 27% of respondents, yet the majority surprisingly judged such evidence-based websites as untrustworthy [5]. The ability to judge the trustworthiness of health information depends on an adequate level of health literacy. Yet, the latest Health Literacy Survey in Europe (HLS-EU) in 2020 showed that at least 1 out of 10 participants (12.4%) had inadequate health literacy, and almost every second respondent (47.6%) in the total sample had limited (inadequate or problematic) health literacy, with substantial differences between member states [6].

Health literacy is a multifaceted construct that "entails people's knowledge, motivation and competences to access, understand, appraise and apply health information in order to make judgments and take decisions in every-day life concerning health care, disease prevention and health promotion" [7].

Low health literacy is known to negatively affect health [8]; it leads to poor health decisions and inferior treatment outcomes as well as higher hospitalization, morbidity, and premature death rates [7]. Low health literacy may also lead to disadvantageous consumer choices when confronted with exaggerated claims about health services, over-the-counter medications, or nutritional supplements. It has been shown that people with lower health literacy have greater difficulty evaluating and differentiating low-quality from high-quality health information [9].

Furthermore, strengthening health literacy is also an important factor for achieving the United Nation's Sustainable Development Goal Number 3: ensuring healthy lives and promoting well-being for all at all ages [10].

Critical health literacy, an essential aspect of health literacy, can play a crucial role in addressing these issues. Critical health literacy is a higher order cognitive process that can be defined as "the ability to access, understand and manage health information as well as the ability to assess its credibility and to critically analyze and where appropriate challenge the information" [11]. It leads to individual and community empowerment by raising critical consciousness, improving quality of life and health behavior, and enhancing health-related social and political action [11]. One way to enhance critical health literacy involves employing a tool for assessing the reliability of health information. To be useful, a tool must be based on operationalized quality criteria, should be validated, and must be easily understood and applied by laypersons [12; 13]. According to a 2020 systematic review [13], none of the many preexisting checklist tools fulfill all these requirements. This is also true for the recently developed Mapping the Quality of Health Information (MAPPinfo) Checklist [12]. Although it is a validated tool based on operationalized and evidence-based criteria, untrained laypersons may find it too complex to apply in their daily routines.

The aim of our study was, therefore, to develop a validated checklist for laypersons to evaluate the trustworthiness of online health information without prior training.

Methods

We employed a multistage mixed-methods approach to develop our checklist. In short, the development process included a comprehensive literature review including extracting possible checklist items, an expert Delphi study, cognitive interviews with lay users, and application tests with laypersons and the research team using a webpage test set to determine the suitability of potential checklist items for predicting the factual correctness of health information webpage content (see Figure 1 for an overview of the development process).

We developed a protocol a priori and published and registered it in the Open Science Framework

(OSF) retrospectively [14]. During the study, we made a few amendments to the protocol. To ensure comprehensive reporting, we adhered to the Standards for Reporting Qualitative Research (SRQR) [15] and the American Psychological Association Style Journal Article Reporting Standards for mixed-methods research [16].

Working definitions

Online health information

We defined *online health information* as any health-related information that a person will find on a webpage when using an online search engine (such as Google).

Trustworthiness of online health information

For the purpose of this study, we define *trustworthiness of online health information* as the degree to which the information on a webpage can be trusted by laypersons to be reliable and valid. This definition aligns with Viviani and Pasi's [17] conceptualization of trustworthiness as a key dimension of credibility, focusing specifically on the audience's perception. We use this as a proxy for the correctness, that is, objective accuracy of the information.

Laypersons

Following the Cambridge Dictionary [18], we have defined *layperson* as someone who is not an expert in or does not have a detailed knowledge of health-related topics, that is, persons without a formal health-related education (i.e., no medical doctors, nurses, or other health care professionals or health researchers, etc.).

Literature search and selection

In a first step, we carried out a literature review for existing tools and/or checklists for the assessment or evaluation of (online) health information as well as for quality criteria and indicators for trustworthy and reliable health information.

While we did not conduct a systematic review, we aimed to identify a broad variety of tools available. Exploratory searches revealed that many relevant tools were grey literature, so we adopted a search process inspired by the Tailored Approach by Cooper et al. [19], which combined a variety of search methods:

- An information specialist (IK) conducted precision-focused searches for reviews about the evaluation of health information in PubMed, Scopus.com, Epistemonikos.org, and Library, Information Science & Technology Abstracts (searched via Ebsco).
- Several team members (IK, BK, FKA, CK, BL, EK, UG) carried out independent web searches using Google, Google Scholar, and the Bielefeld Academic Search Engine (BASE).
- The reference lists of all identified (systematic) reviews were checked for citations of eligible tools or publications.

All searches were conducted between March and September 2021. Search results were collected in an Endnote 20 Library and checked for duplicates.

One person screened all the results, and a second person checked the selected documents for completeness. We included published and grey literature documents in English or German language (i.e., journal articles, reports, webpages, preprints) from the following categories:

- Checklists and tools for the evaluation of online health and other online information aimed at laypersons or professionals
- Research on the concepts, quality criteria, and indicators relevant for the evaluation of online health information

- Systematic reviews of the tools/checklists or concepts/quality criteria/indicators

We excluded literature that either exclusively described the evaluation of the layout, design, comprehensibility, or readability of (online) health information as well as literature applicable only for evaluating health information for specific diseases and point-of-care information.

Data extraction

We developed a data extraction matrix in Microsoft Excel to facilitate the data extraction and item generation step. The data extraction consisted of three parts: 1) general information from the included references, that is, reference details, reference type, checklist/tool name, target group(s), application field, and validation status; 2) the names of domains and subdomains/subcategories, literal item text and original answer categories (if available), or description of the domain/criterion/indicator (if applicable); 3) for each included (systematic) review on tools/checklists, we extracted the included names and references of the tool/checklist and added the most frequently mentioned tools/checklists that we had not included yet.

After the first few data extractions, we discussed unclarities within the project team and revised the data extraction form accordingly. If references had to be excluded during the data extraction, these decisions were always discussed with another research team member.

Categorization of items

The next step after the data extraction was the categorization of items into content categories. One research team member developed a draft categorization scheme inductively, which we discussed as a team. We developed a categorization scheme iteratively including the names of the overall categories and subcategories, a description of each subcategory, and coding examples (see Table 1).

We excluded unsuitable criteria and marked them as "excludes" based on adapted feasibility criteria by Provost 2006 [20], which we defined a priori:

- Timeliness: Criterion must be assessable in a short time.
- Expertise independent: Criterion must be assessable without prior topic-specific knowledge.
- Externability: Criterion must be assessable without using secondary sources.
- Generalizability: Criterion must be applicable to all possible (online) health information (i.e., not focused on specific aspects of a disease or a specific health topic).

We added further subcategories for the "excludes" domain during the process:

- Scope: Content of the item is not within our scope.
- Overall category: Only the item category is specified, no items.
- Unclear: Item is too unclear to be used in a checklist.
- Open question: it is not possible to formulate a yes/no question.

Table 1: Item categorization by category

Subcategory	**Description**	**Example**	**Number of items after categorization step (n=1740)**	**Number of items after deduplication and merging (n=449)**	**Number of items for the 1st Delphi round (n=46)**
Functional/ technical aspects			**163 (9%)**	**25 (6%)**	**NA**[a]
F.0 General technical aspects	A broad category if the item is generally about technology or several technical aspects are covered at once		9	0	NA
F.1 Technical	Technical accessibility of the	Browser	40	10	NA

Subcategory	Description	Example	Number of items after categorization step (n=1740)	Number of items after deduplication and merging (n=449)	Number of items for the 1st Delphi round (n=46)
accessibility	website	compatibility, website loading times, registration			
F.2 Navigation	Navigation through the website	Site menu, simple navigation, internal search engine	55	5	NA
F.3 Interactivity	Interactive offers on the site	Direct feedback options on the site (e.g., comment field, rating options, chat room, forum, etc.), download options	38	3	NA
F.4 Accessibility/ personalization	Possibility of customization (of the presentation) to individual needs	Customization of font sizes/types, text-only option, several languages available	17	3	NA
F.5 Other (technology)	An item belongs in this category but does not fit into any of the available subcategories		4	1	NA
Transparency			**450 (26%)**	**199 (44%)**	**9 (20%)**
T.0 General transparency	A category if the item is generally about transparency and cannot be assigned to any of the subcategories (or must be assigned to several)		82	0	0
T.1 Up-to-date information	Information on the content's actuality is available	Date of creation and/or last update stated; update frequency stated	82	31	2
T.2 Author and copyright	Identification of the content authorship/ creator, copyright information	Name of the author/authorship, email address, imprint	81	43	2
T.3 Author information	Further information that provides an indication of the content creator's (author's) context	Mention of the content creator's role, profession, education, name, author qualifications; mention of affiliation, (type of) institution	52	29	0
T.4 Disclosure and funding	Identification of funding sources, cooperation partners, conflicts of interest, identification of advertising	Naming of sponsors, financing institution, adverts/advertising is shown, separation of editorial and advertising content	126	47	4
T.5 Standards/certificates	Information on whether the content creators adhere to certain external guidelines or whether the source is externally certified	HONcode Standard	6	4	1
T.6 Privacy and data protection	Information/functions regarding privacy and data protection	General disclaimer, information on the use of cookies, message when leaving a secure website	46	24	0
T.7 Target group	Definition of the website's target group	Explicit explanation of who the website and its content are	17	5	0

Subcategory	**Description**	**Example**	**Number of items after categori-zation step (n=1740)**	**Number of items after dedupli-cation and merging (n=449)**	**Number of items for the 1st Delphi round (n=46)**
		aimed at			
T.8 Purpose	Explicit mention of the website's purpose and goals	Educational purposes, commercial interests (this does not refer to subjective assumptions as to what the purpose might be)	40	16	0
T.9 Other (transparency)	An item belongs in this category but does not fit into any of the available subcategories		0	0	0
Presentation of information			**402 (23%)**	**66 (15%)**	**22 (48%)**
I.0 General presentation of information	A category if the item is generally about the presentation of information and cannot be assigned to any of the subcategories		7	0	0
I.1 Balance	Presentation of information is one-sided or different aspects are highlighted	Complete, scientifically balanced comparison of information; in addition to benefits, risks, possible complications, consequences of no treatment are mentioned, other treatment options are listed; information and recommendations are clearly distinct	215	34	12
I.2 References	Details on the origin of the information provided are available	Availability of information sources, references, links to sources	66	9	3
I.3 Level of evidence	Information on the categorization of the significance/evidence level/quality of the information is available	The text explicitly addresses how well or poorly the findings/data are scientifically validated. Gaps in knowledge are addressed	30	10	2
I.4 Quality assurance	Presence of quality assurance processes	It is clear to users that there is an editorial review process and how it works	21	4	1
I.5 Methods	The procedure for creating the content is described	Transparent systematic literature search, evaluation according to the GRADE criteria	22	4	3
I.6 Further information	Indication of further websites, literature, contacts	Links to further content; references to offers of help; disclaimer: note that health information does not replace a	41	5	1

Subcategory	**Description**	**Example**	**Number of items after categori-zation step (n=1740)**	**Number of items after dedupli-cation and merging (n=449)**	**Number of items for the 1st Delphi round (n=46)**
		doctor visit			
I.7 Other (presentation of information)	(Not meant are cited sources or the author's contact information)		0	0	0
Presentation (linguistic and visual)			**303 (17%)**	**99 (22%)**	**7 (15%)**
D.0 General presentation	A category if the item is generally about visual and linguistic presentation or if several subcriteria are covered at once		9	2	0
D.1 Comprehensibility	Appropriate presentation of information for the target group (content and form)	Appropriateness of language, sentence structure, readability, comprehensible presentation of figures, explanation of technical terms, degree of complexity is adapted to the target group	116	42	1
D.2 Layout	The website and content's layout and design	The content's layout (e.g., font, font size, structure) and visual presentation (e.g., illustrations, graphics, tables), clarity (e.g., short text), appropriate color contrast, appealing design, illustration quality, graphics, tables (readability), etc.	126	40	1
D.3 Linguistic style	Type of information preparation (neutral, emotional, sensationalist)	Objective, neutral language style, no judgmental, emotionalizing language	42	14	5
D.4 Formal correctness	Reference to the care taken in content creation: Are there formal errors in texts or presentation?	Correct grammar and spelling, figures in text and tables match	10	1	0
D.5 Other (presentation)	An item belongs in this category but does not fit into any of the available subcategories		0	0	0
User perception			**185 (11%)**	**60 (13%)**	**8 (17%)**
N.0 General user perception	A category if the item is generally about user perception and cannot be assigned to any of the subcategories		10	2	0
N.1 Emotion	What feelings does the health information trigger?	Subjective assessment based on the formal and content-related presentation of the information	7	4	1
N.2 Familiarity and	Is the source already known?	Subjective assessment	40	13	1

Subcategory	Description	Example	Number of items after categorization step (n=1740)	Number of items after deduplication and merging (n=449)	Number of items for the 1st Delphi round (n=46)
reputation	How are the source's and site creator's reputation assessed?	based on the content but also on external factors (prior knowledge, bias, background, etc.)			
N.3 Trustworthiness of content	Individual assessment of the content's trustworthiness	Subjective assessment based on the content but also on external factors (prior knowledge, bias, background, etc.)	47	15	3
N.4 Trustworthiness of references	Individual assessment of the trustworthiness of the sources and of further links	Subjective assessment of the trustworthiness of specified sources of information and data, of further literature, of references to offers of help	10	5	1
N.5 Trustworthiness in general	General assessment of the trustworthiness of the entire website/the entire offer	Trust in the organization creating the offer, assessment of the website name or URL	29	13	2
N.6 Relevance/usefulness	Individual assessment of the content's usefulness/practical applicability	Influence of the information on individual user behavior; relevance to the user's original need for information, specific examples for users are given	42	8	0
N.7 Other (user perception)	An item belongs in this category but does not fit into any of the available subcategories		0	0	0
Excludes			**237 (14%)**	**NA**[b]	**NA**[b]
X_external source	Can only be assessed by using additional tools/sources	If the institution's or author's reputation is unknown and further internet research would be necessary	36	NA	NA
X_Scope	The item's content is not within this study's scope	Items on laypersons' search behavior, items for health information producers	82	NA	NA
X_Generalizability	The item cannot be generalized, but can only be used for a specific type of health information	The item is only relevant for cancer information; the item refers to a comparison of treatment methods and may not be used for health information on the prevalence of diseases	59	NA	NA

[a] items from the functional/technical aspects category were excluded as they were deemed not useful for the checklist

[b] excludes were not further deduplicated or translated

Abbreviations: GRADE, Grading of Recommendations, Assessment, Development and Evaluation; n, number; NA, not applicable

We piloted the categorization process and revised the category system. For each category, there was one subcategory called “other,” where we added items provisionally and either created new subcategories or decided on a subcategory after joint discussion. We also adapted subcategories for the “excludes” (see Table 1).

Four members of the research team (CK, EK, IK, UG) categorized all the available items, and a second person checked the categorization. Any uncertainties were discussed together.

Condensation of items

The first step of condensing the item list was the merging of duplicate and similarly worded items as well as the translation of English items into German. As the literal items were very diverse, we also unified their wording in two ways: first, we changed the wording so that each item was a question and, second, we determined that the answer “yes” always indicates a trustworthy text. This was done in one step by one person. We divided all the items among four team researchers (CK, EK, IK, UG) and developed a procedure for merging and item formulation to ensure that the process was done similarly by each. We noted discussion points and discussed unclarities and uncertainties during the condensation process among the team. After the first round of merging and item formulation, the process was checked by a second person, and discrepancies or conflicts were resolved by discussion.

After merging the duplicates, we still had too many items to conduct a meaningful expert Delphi study. We therefore applied a further reduction step. For this, we created an Excel file matrix with the item text, the category and subcategory allocation, and two rating options on a scale from 1 to 5, with 1 being “very poor appropriateness/applicability” and 5 being “very good appropriateness/ applicability:” one for item’s appropriateness (i.e., how appropriate is the item for assessing whether the health information is trustworthy) and one for the item’s applicability (i.e., how easily can the item be applied by checklist users without prior knowledge). After being given detailed instructions, eight members of the research group (EK, FKA, BL, UG, BK, CK, IK, IS) determined the ratings separately. We calculated each item’s mean rating to decide which to keep for the expert Delphi study. Items that scored very low (i.e., achieving only ratings of 1 and 2 from all eight research group members) were eliminated without further inspection. All other items within each subcategory were checked to determine which was the highest rated alternative; the other items were eliminated. The final selection was made according to content-related aspects and while accounting for the target number of approximately 50 items for the expert Delphi study.

Expert Delphi study

Next, we conducted an expert Delphi study with the aim of eliciting the opinions of invited experts and establishing a consensus on the most pertinent items for the checklist [21; 22]. Our objective was to assemble a diverse panel of experts specializing in patient communication, health information, and general online communication. To identify potential experts, we searched the internet, utilized established contacts, took suggestions from within the research team, and applied a snowballing approach. Out of the twelve experts invited, six participated in the expert Delphi study.

During an initial online meeting with the invited experts in May 2022, we clarified the research objectives and study process and presented the categories and subcategories of the preliminary item checklist. The experts received a comprehensive information sheet as well as instructions for participating in the Delphi study, using the same Excel file matrix layout as in the previous item condensation step. In addition to rating the items from 1 to 5 for appropriateness and applicability, the experts were given the opportunity to suggest improvements in the wording of existing items or to provide comments. They were also encouraged to propose up to five new items that they felt were missing. We calculated the mean, standard deviation, and minimum and maximum rating values for each item and incorporated suggestions for improvements and ideas for missing items. In the second Delphi round, the expert panel rated the items again for appropriateness and applicability. Descriptive statistics were calculated again, and items were selected based on the experts' scores and on content considerations. The result of the two-step Delphi study was an interim checklist version 1.

Cognitive interviews

Next, the first version of the interim checklist was presented to laypersons in a cognitive interview. The cognitive interview methodology examines individuals' cognitive processes when processing information and responding to questions [23] and is widely used in questionnaire pretesting [24; 25]. Cognitive interviewing was used in this study to identify problems with item comprehension and to evaluate the usability and applicability of interim checklist version 1. We used both the think-aloud and verbal probing techniques [26; 27], where participants were explicitly instructed to verbalize their thought processes during the cognitive interview.

Participant selection and sample

The sample size and characteristics were based on the recommendations of the cognitive interview [28]. We used the purposive sampling method and maximum variance strategy for participant recruitment and selection [29]. Recruitment strategies included advertising the study through social media, using established contacts with self-help groups, and drawing on our personal networks. Interested individuals were asked to provide basic sociodemographic information to populate our sampling grid. We selected participants to reflect the current sociodemographic pattern in terms of age groups, gender, educational level, and migrant background. Eligible participants were adults of any gender, 18 years and older, with a basic education level (i.e., compulsory schooling or high school diploma) and no formal health-related education (i.e., no medical doctors, nurses, or other health care professionals or health researchers, etc.), who regularly use the internet to seek answers to health-related questions (i.e., at least once in the last three months). Our target group is fluent in German with no cognitive limitations. Before the interview, study participants were given an information sheet explaining the study purpose, privacy protection policy, and cognitive interview procedure.

Data collection and ethical considerations

Interviews were conducted either online via MS Teams (n=16) or face-to-face (n=3), according to the study participants' preferences. Participants provided their informed consent verbally prior to starting the interviews and received a gift voucher of 30 Euro afterward to compensate for their time and effort. The study was conducted in accordance with art. 13 GDPR of the University for Continuing Education Krems' privacy policy and was approved by the university's Ethics Commission under reference number EK GZ 23/2021-2024.

We developed an interview guide with probing questions prior to the first interview and adapted it after discussion among the research team. Furthermore, we prepared a version of interim checklist version 1 with a visually appealing layout along with brief written instructions for use. Interviews were conducted by three researchers (EK, CK, UG), recorded using MS Teams or a recording device, and transcribed verbally. The interviewers preselected three sample health information webpages for the cognitive test. A few days before the interview, study participants received an email from the interviewer with a link to the health information webpage and an invitation to read the text in advance. While responding to the checklist items, participants were free to view the sample health information webpage and to surf to webpages beyond the initial page. Participants were asked to use the checklist on the provided health information webpages. They were prompted to comment on any potentially problematic words or phrases in the checklist and whether they had any problems answering the checklist items. Additionally, probing questions were used to identify or explore potential sources of response errors. Finally, study participants were asked to evaluate the checklist's layout, usability, and overall quality. The interviews lasted between 33 and 91 minutes (mean: 56 minutes). Further, each interviewer completed a structured protocol on their observations of the problems encountered, comments on the nature of the problems, or whether there was no evidence of a problem.

Data analysis

We employed the framework approach [30] to analyze the interview data. The three interviewing researchers (UG, CK, EK) started by reading the transcripts of the interviews they conducted and their notes on the observation sheet. They marked potential problems (e.g., problems with comprehension, knowledge, or the instructions or checklist layout) and extracted the data into a structured matrix. Any reasons for unclarities, suggestions for revision, or comments on which words and phrases were particularly well understood were extracted into the structured data analysis table. The responses to the additional questions were also paraphrased in the matrix. Each interviewer extracted the data from their own interviews, and the overall assessment was done in several discussion rounds among all three analyzers (UG, CK, EK). Questionable or controversial items were identified and suggestions for any revisions discussed among the whole research team. The cognitive testing was done in two rounds: after approximately half of the interviews, the checklist was adapted, the order of the checklist items was rearranged, and explanatory item subtitles were added. The analysis of all the cognitive interviews and discussions among the whole research team resulted in interim checklist version 2, which was used as the basis for the application testing with additional lay users.

Application testing with lay users

We tested the applicability of interim checklist version 2 with another set of potential future lay users.

We recruited 20 additional laypersons, including interested individuals not previously selected for the cognitive interviews. Participants were asked to apply the checklist to 15 selected health information webpages. After returning the completed checklist documents, they received a link for a short online post-task questionnaire. The 15 health information webpages were a purposively selected sample from a test set consisting of 100 health information webpages (see next chapter) representing a range of different types of health information—around half of which provided the information and underlying evidence correctly and half of which did not. The post-task questionnaire asked participants to rate how easy or difficult they found using the checklist on a scale from 1 (very easy) to 10 (very difficult). It also provided a list of all 23 items from interim checklist version 2, where participants could mark those that were difficult to answer and provide an explanation for why. Furthermore, we asked participants to mark the items that they perceived as particularly important for evaluating health information. Participants were also given the opportunity to suggest the rephrasing of item texts. The interim checklist version 2 items used in the application test can be found in Table 4.

Each participant was remunerated for their effort with a voucher of 225 Euro after having returned the completed checklists and questionnaire.

We considered the quantitative and qualitative results from the application testing with lay users in the final selection of items for the final checklist. We calculated Fleiss' Kappa as an interrater reliability measure for the laypersons' ratings for each item to see whether they arrived at very different answers. The aim was to sort out such items because they are obviously unsuitable for an objective assessment of health information webpage by lay users. The data from the post-task questionnaires was analyzed in two ways: the quantitative data was analyzed descriptively, and the qualitative data in the open answers by means of content analysis.

Application testing with research team members

Interim checklist version 2 was also applied to the full test set of 100 health information webpages by members of the research team (BL, IK, DL, BK, IM, UG). The health information was considered trustworthy if between 16 and 23 out of a total of 23 items were answered with “yes.” To maximize the objectivity, each webpage was assessed independently by two researchers, and disagreements were resolved by discussion or consultation with a third person. For each item, agreement among the research team (before reaching a consented rating) ranged from 66 to 97 (out of a possible 100 webpages from the test set). We used the quantitative data from the checklists (i.e., the item ratings and interrater agreement) for statistical analysis (see chapter “Statistical analysis for predictive validity”).

In addition, after the application test, each research team member took notes on the specific item’s difficulty, importance, or possible ambiguity. These qualitative results were also considered in the overall discussion and final selection of items. We also calculated the mean difference between the consented expert ratings and each layperson’s rating and took this into account for the final item selection.

Operationalization of health information trustworthiness

We aimed to assess each item’s suitability in predicting health information trustworthiness. To accomplish this, we operationalized trustworthiness as the factual correctness of the answer to a given health question on a health information webpage.

To test the predictive power of our items, we constructed a test set of 100 health information webpages that provided answers to ten common health questions such as, for example, does vitamin C help with a cold? Does arthroscopy help with osteoarthritis? Or do omega-3 fatty acids prevent cardiovascular diseases? (see Table A4 in the supplement for a full list). For each of these questions, we conducted a Google search to find ten webpages covering the respective topics. We tried to balance the factual correctness, with approximately half of the webpages accurately reflecting the evidence base of the respective health question, while the other half did not.

The evidence base behind these ten health questions had been researched within the previous year by a team of specially trained science journalists from the fact-checking platform “Medizin transparent” [31], a project by Cochrane Austria and the Department for Evidence-based Medicine and Evaluation at the University for Continuing Education Krems, Austria. Medizin transparent is also a certified signatory of the International Fact Checking Network at the Poynter Institute [32]. It applies an ultra-rapid evidence synthesis method [33] starting with a systematic literature search followed by a critical appraisal of the studies’ risk of bias. The strength of the evidence is then evaluated using the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) approach [34] and is reflected by four categories: very low, low, moderate, and high strength of evidence. For the use of Medizin transparent, we expanded this scale to include a directional dimension indicating whether the evidence supports a treatment effect or no treatment effect. The result was a seven-level scale ranging from -3 (indicating high strength of evidence that the treatment is ineffective) to 3 (indicating high strength of evidence that the treatment is effective), with a midpoint of 0 indicating insufficient evidence for or against the treatment effect.

For the purpose of this study, we defined an analogous seven-level scale to express the *claimed* strength of evidence implied by each of the 100 health information webpages while accounting for the accepted terminology for the different GRADE categories [35]:

-3: The webpage *does not mention any doubts* about the ineffectiveness of the treatment.
-2: The webpage implies that the treatment is *probably/likely* ineffective.
-1: The webpage implies that the treatment *may or could be* ineffective.

0: The webpage implies that it is unclear or insufficiently researched whether the treatment is effective.
1: The webpage implies that the treatment *may or could be* effective.
2: The webpage implies that the treatment is *probably/likely* effective.
3: The webpage *does not mention any doubts* about the effectiveness of the treatment.

Two members of the research team (BK, IM) independently assessed the *claimed* strength of evidence for each of the 100 health information webpages. All disagreements were solved through discussion to arrive at a final assessment.

As previously described [36], we operationalized the degree of factual correctness as the difference between the *factual* strength of evidence and the *claimed* strength of evidence and interpreted it as a proxy measure of a webpage's trustworthiness. If this difference was more than 1 for a given webpage (out of a possible range between 0 and 6), we defined the webpage as containing incorrect information. We defined the information as correct if the difference was 0 or 1, accounting for a small level of uncertainty in the evidence ratings. However, if the *factual* strength of evidence was 0 (insufficient) for a certain health question, we also defined a *claimed* strength of evidence of 1 or -1 (maybe effective or maybe ineffective) as incorrect. This is because numerous complementary or alternative medicine treatments have been claimed to be potentially effective despite a clear lack of proven efficacy as well as plausibility.

Statistical analysis for predictive validity

The statistical analysis method was designed to answer the question: which items can best predict the factual correctness of health information statements on a webpage? We used the results of our application tests of trustworthiness and factual correctness to calculate each item's predictive validity.

To analyze the relationship between the factual correctness and item ratings across specific items of interest, a Bayesian logistic regression model was employed and specified using the brms package in R (version 4.2.2) [37]. Data wrangling visualization was done via the tidyverse package [38].

The dependent variable, binary, indicated whether a given response was correct (1) or incorrect (0). We used the interaction between the rating (a continuous variable) and ID_Item (a variable representing the individual items) as a predictor. This interaction term allows the effect of rating on the probability of a correct response to vary across different items.

The model was defined as follows:
correct∼1+rating:ID_Item

This formula indicates that the model includes an intercept and an interaction between the rating and ID_Item. The outcome variable was modeled using a Bernoulli distribution. The prior for the intercept was a normal distribution with a mean of 0 and a standard deviation of 2 (normal(0, 2)). The prior for the regression coefficients (b) was also a normal distribution with a mean of 0 and a standard deviation of 2 (normal(0, 2)). These priors reflect weakly informative assumptions, allowing the data to primarily inform the posterior distributions.

The items were finally ordered according to their predictive validity. We chose different sets of items depending on their predictive validity, layperson judgments regarding the difficulty of answering the item, the item's perceived importance, any layperson comments or research team notes, the regression coefficient for the predictive validity, the interrater agreement separately among laypeople and among the research team, and the agreement between laypersons and the research team.

Results

Literature search and selection

Because of the varied nature of the included documents (e.g., published articles, webpages, downloadable PDFs of checklists, tools, and grey literature reports, etc.), we could not adhere to a strict separation of abstract and full-text screening. Each document was screened by one person, with a second person verifying the inclusion/exclusion decision.

The initial search steps (database and web searches) identified 200 documents that were further evaluated, resulting in 52 tools and checklists that were included for the data extraction.

We also identified 11 relevant reviews. From each included systematic or literature review on tools and checklists, we extracted the tools' and checklists' names and references. Out of the resulting 140 additional references, 21 tools or checklists that met our eligibility criteria were added for the data extraction. This led to a final set of 73 included documents.

Data extraction and categorization of items

We performed a data extraction of general information for all 73 included documents. Of these, 36 were journal articles, 27 were webpages, three were books or book sections, three were reports, two were conference papers, and one each was a preprint and an unpublished paper. Of the 73 documents, 46 (63%) were tools or checklists, more than half of which were aimed at laypersons, consumers, or patients (27/46, 59%), approximately a third at experts (13/46, 28%), that is, health information producers or health care professionals, and three (7%) did not have a specific target audience.

We extracted a total of 1,740 items and either categorized all items into five content categories or marked them as excluded. Table 1 shows the subcategories for each category along with a description and examples as well as the number of items for each subcategory. The largest number of items (n=450, 26%) belonged to the transparency category, followed by presentation of information (n=402, 23%). The least number of items belonged to the functional/technical aspects category (n=163, 9%), which is unsurprising as this was not our search's main focus. About 14% (n=237) of the items had to be excluded, either because of the wrong scope, the need for prior knowledge, the lack of generalizability, or the need to use an external source to apply the item, all of which we defined a priori as reasons for exclusion.

Condensation of items

The first step in condensing the item list resulted in a reduction from 1,740 to 449 items. Table 1 shows the distribution of items among the respective content categories and subcategories. A further reduction step using the research team members' ratings of the items' appropriateness and applicability resulted in a list of 46 items (see last column in Table 1 for the content distribution).

Expert Delphi study

The descriptive results of the ratings of the 46 items, grouped according to the subcategories, in the first Delphi round can be found in the Appendix Table A1. The experts, two from Germany and four from Austria, four female and two male, made some minor suggestions for improving the wording of the items. A major revision according to the experts' feedback was to formulate statements instead of questions. Overall, the participating experts suggested including six additional items. In Delphi round 2, the experts were given a total of 52 items, both the original 46 items with new wording suggestions along with the results from round 1, plus six additional items. The descriptive results of the ratings of the 52 items in the second Delphi round can be found in the Appendix Table A2. Overall, the applicability ratings changed more than the appropriateness ratings during the two Delphi rounds (see Appendix Table A3).

After the Delphi results were available, the research team convened to discuss the results and perform the item selection. For the selection process, we did not have a specified a priori selection criteria as we assumed that the decisions could only be made after a thorough discussion of the results. In general, the experts gave higher ratings for appropriateness (mean min and max values: 3.2 and 5) than for applicability (mean min and max values: 2.2 and 4.8). We chose a stepwise process for the selection and first excluded items with a mean rating of 4.0 or lower on the appropriateness scale and 3.0 or lower on the applicability scale. Then we excluded items with an applicability rating between 3.0 and 4.0. We carefully scrutinized the remaining list and cross-checked with the excluded items that all content categories were still represented and included the best-rated items in case no item for the category was present. For example, we included an item with an applicability rating of 3.17 because we considered the topic of quality assurance as very important, and it was only reflected in one item. We also cross-checked once again for any redundancies.

The selection process resulted in interim checklist version 1 consisting of 23 items. In general, the items selected received higher ratings on both scales on average than the items not selected (Table 2).

Table 2: Ratings in Delphi Round 2 for each item subcategory, subgroups for items selected/not selected

Subcategory	**Items not selected**			**Items selected**		
	# Items	**Appr. M**	**Appl. M**	**# Items**	**Appr. M**	**Appl. M**
D.1 Comprehensibility	2	3.52	3.27	1	4.00	4.50
D.2 Layout	1	3.67	4.17	-	-	-
D.3 Linguistic style	3	4.17	3.50	1	4.33	4.17
I.1. Balance	9	4.35	3.35	7	4.53	4.33
I.2 References	2	4.58	2.83	1	4.67	3.67
I.3 Level of evidence	1	4.67	3.67	1	4.83	3.67
I.4 Quality assurance	-	-	-	1	4.00	3.17
I.5 Methods	1	4.50	3.17	2	4.58	3.67
I.6 Further information	-	-	-	1	4.33	4.67
N.1 Emotion	-	-	-	1	4.50	4.17
N.2 Familiarity and reputation	1	4.17	2.83	-	-	-
N.3 Trustworthiness of content	2	4.17	3.25	1	4.00	4.00
N.4 Trustworthiness of references	1	5.00	3.67	-	-	-
N.5 Trustworthiness in general	1	4.17	3.83	1	4.00	4.00
T.1 Up-to-date information	1	4.50	4.33	1	5.00	4.83
T.2 Author and copyright	1	4.67	3.83	2	4.67	4.42
T.4 Disclosure and funding	3	4.56	3.50	1	4.83	4.17
T.5 Standards/ certificates	-	-	-	1	4.33	4.33

Abbreviations: Appr., appropriateness; Appl., applicability; M, mean; T, transparency; I, presentation of information; D, presentation (linguistic and visual); N, user perception

Cognitive interviews

We conducted two rounds of cognitive interviews, involving nine persons in the first round and ten in the second. We managed to recruit a diverse set of persons for inclusion, with 58% (n=11) female, 58% (n=11) between 40 and 64 years old, 11% (n=2) and 26% (n=5) with a lower educational background of either compulsory school only or apprenticeship/vocational secondary school, respectively, and 21% (n=4) with a migration background (Table 3).

Table 3: Characteristics of lay participants in cognitive interviews and application tests

	Cognitive interviews (n=19)	**Application tests (n=20)**
Gender		
female	11 (58%)	12 (60%)
male	8 (42%)	8 (40%)
Age groups in years		
19–39	5[a] (26%)	9 (45%)
40–64	11 (58%)	10 (50%)
65–84	3 (16%)	1 (5%)
Highest education level		
Compulsory school	2 (11%)	1 (5%)
Apprenticeship/vocational secondary school	5 (26%)	10 (50%)
Upper secondary school/vocational upper secondary school	8 (42%)	8 (40%)
College/academy/university	4 (21%)	1 (5%)
Migration background		
no	15 (79%)	20 (100%)
yes	4 (Germany, Turkey, Hungary, USA) (21%)	-

[a] one participant was 16 years old
Abbreviation: n, number

The first round results showed that many words used in interim checklist version 1 were difficult to understand for the users, and some items were rated as less important to the lay users than to the experts. Therefore, we made large modifications to the checklist: We reworded 17 of the 23 items, deleted one item, and added another; we revised the instructions on how to interpret the checklist results; we reordered the items in a way in which the information on a webpage would usually be presented to avoid too much scrolling by the users.

In the second round, further modifications to interim checklist version 1 were implemented, for example, adding subtitles to structure the checklist and provide more guidance to the users. Two additional items were tested with the last seven cognitive interviews.

After analysis and discussions among the whole research team, a second interim checklist version was created for use in the application test (see Table 4). The wording of thirteen items was changed or text was added to the item to give tips on where to find the respective information on the webpage. Furthermore, the order of the items was changed once more.

Already in the first round, lay users made remarks on single items that were less important to them than expert users, for example, “It is clear to which target group the health information applies.” As this was an item deemed important by the experts in the Delphi study, we kept it for the application test but added an example because the term *target group* was ambiguous for some users. In general, many of the suggested modifications led to shorter and more straightforward sentences and the explicit use of examples (i.e., “for example”). We also removed redundancies and used words that were unlikely to be misunderstood, and we avoided negative constructions as these are difficult to understand in German. During the analysis, we realized that some items may be more important for health information creators than for health information consumers, for example, one item that we dropped after the cognitive interviews: “The health information mentions how well the differences between men and women have been researched.”

Table 4: Items of interim checklist version 2 with the lay application test results

	Item text	**Items perceived as difficult to answer by lay users (n=20)**	**Items perceived as important by lay users (n=20)**	**Regression coefficient, error, 96% CI[a]**	**Agreement among lay users (Fleiss’ Kappa)**	**Agreement among research team members (100=complete agreement)[b]**	**Item part of the final checklist with wording**
1	The title or subtitle is factual and neutral.	1	0	0.56, 0.24, (0.07, 1.02)	0.30	80	
2	The health information does not contain advertising related to the health problem.	1	4	0.46, 0.25, (-0.02, 0.94)	0.52	85	
3	The health information does not contain any advertising.	3	4	0.82, 0.31, (0.22, 1.44)	0.50	84	The health information does not contain advertising.
4	The health information is from an independent institution that is unlikely to make money from others’ health. For example, not from a company that sells medicines or food supplements. *** **Tip**: Such information is often found in	6	14 (70%)	0.67, 0.24, (0.2, 1.14)	0.48	81	The health information is from an independent institution that presumably does not make any money from our health (e.g., no suppliers of medicines or food supplements. .. .).

	Item text	Items perceived as difficult to answer by lay users (n=20)	Items perceived as important by lay users (n=20)	Regression coefficient, error, 96% CI[a]	Agreement among lay users (Fleiss' Kappa)	Agreement among research team members (100=complete agreement)[b]	Item part of the final checklist with wording
	the imprint.						
5	There is a quality seal on the website. Examples of reliable quality seals are: afgis logo (Action Forum Health Information System). International Fact Checking Network.	5	1	1.16, 0.45, (0.27, 2.06)	0.39	94	
6	I feel the information is presented in a balanced way. For example, different treatment options are described. Or besides positive effects, side effects or disadvantages are also described.	1	11 (55%)	1.12, 0.28, (0.57, 1.66)	0.30	72	I feel the information is presented in a balanced way (the health information describes e.g., advantages and disadvantages, different treatment options. ...).
7	The health information describes treatment effects that I can feel myself. For example, "the feeling of dizziness decreases" instead of "blood pressure decreases."	6	1	0.48, 0.23, (0.04, 0.93)	0.16	66	
8	The health information also tells me if and what consequences there are if I	8 (40%)	1	0.52, 0.47, (-0.42, 1.4)	0.08	74	

	Item text	Items perceived as difficult to answer by lay users (n=20)	Items perceived as important by lay users (n=20)	Regression coefficient, error, 96% CI[a]	Agreement among lay users (Fleiss' Kappa)	Agreement among research team members (100=complete agreement)[b]	Item part of the final checklist with wording
	choose not to have a treatment.	🟥					
9	The health information mentions whether there are differences between men and women. For example, differences in a treatment's symptoms, effects, or side effects.	4	2	0.35, 0.45, (-0.53, 1.21)	0.22	87	
10	The health information does not make exaggerated or sensational statements.	1	1	0.5, 0.25, (0.01, 0.97)	0.28	72	
11	The language is factual and neutral.	0	5	0.43, 0.24, (-0.05, 0.9)	0.12	72	
12	Technical terms are used sparingly, and their meanings are explained.	0	6	0.67, 0.26, (0.16, 1.2)	0.23	78	Technical terms are used sparingly, and their meaning is explained.
13	It is clear to which target group the health information applies. For example, for men/women, people with certain conditions, etc.	3	1	0.24, 0.23, (-0.23, 0.68)	0.09	76	
14	The health information indicates how well a fact is scientifically proven.	7	11 (55%)	0.87, 0.24, (0.4, 1.34)	0.30	80	The health information indicates how well the facts claimed are scientifically supported.
15	The health information	2	7	0.69, 0.27, (0.16, 1.2)	0.62	85	The health information

	Item text	Items perceived as difficult to answer by lay users (n=20)	Items perceived as important by lay users (n=20)	Regression coefficient, error, 96% CI[a]	Agreement among lay users (Fleiss' Kappa)	Agreement among research team members (100=complete agreement)[b]	Item part of the final checklist with wording
	provides detailed references for the facts mentioned. For example, a list of sources or links to the studies mentioned.						states in detail which sources are behind the facts mentioned (bibliography, links to studies. ...).
16	The health information clearly states that only a doctor can assess my health problem.	0	4	0.7, 0.3, (0.12, 1.29)	0.41	75	
17	After reading the health information, I feel I can make a decision without pressure.	4	1	0.53, 0.24, (0.06, 1.01)	0.10	67	
18	It is clear when the health information was created or updated.	3	4	0.68, 0.24, (0.21, 1.16)	0.67	94	It is clear when the health information was created or updated.
19	It is clear that the health information is up to date. For example, the date it was created or updated.	4	4	0.7, 0.26, (0.19, 1.19)	0.39	88	
20	It is clear that the health information was written by a person or team with suitable scientific training. For example, medical studies or	5	9 (45%)	0.62, 0.34, (-0.06, 1.28)	0.34	86	

	Item text	Items perceived as difficult to answer by lay users (n=20)	Items perceived as important by lay users (n=20)	Regression coefficient, error, 96% CI[a]	Agreement among lay users (Fleiss' Kappa)	Agreement among research team members (100=complete agreement)[b]	Item part of the final checklist with wording
	other health-related training (nursing, pharmacy, biology, etc.). *** **Tip**: Such information is often found on the "About Us" page or in further links.		🟩				
21	It is clear that the health information has been checked by a person with appropriate scientific training. For example, medical studies or other health-related training (nursing, pharmacy, biology, etc.). *** **Tip**: Such information is often found on the "About Us" page or in further links.	9 (45%)	6	0.43, 0.45, (-0.48, 1.3)	0.31	86	
22	It is described how the information was created. For example, which studies were considered and why, and which were not. *** **Tip**: Such information is often found on the "About Us" page or in	8 (40%)	1	1.49, 0.52, (0.49, 2.53)	0.22	85	

	Item text	Items perceived as difficult to answer by lay users (n=20)	Items perceived as important by lay users (n=20)	Regression coefficient, error, 96% CI[a]	Agreement among lay users (Fleiss' Kappa)	Agreement among research team members (100=complete agreement)[b]	Item part of the final checklist with wording
	further links.						
23	The health information states whether and how readers were involved in creating the information. *** **Tip**: Such information is often found on the "About Us" page or in further links.	11 (55%)	0	3.39, 1.07, (1.57, 5.7)[c]	0.02	97	

[a] higher values denote better predictive validity

[b] before the disagreements were solved through discussion or by involving a third person

[c] this item was only ticked in 8 health information webpages; therefore the practical value of the predictive validity is low

red marking: items perceived as difficult by 40% of lay users or more; green marking: items perceived as important by 45% of lay users or more

Abbreviations: n, number; CI, confidence interval

Application test with lay users

We conducted application tests with 20 further lay users (see participant characteristics in Table 3). All 20 lay users returned quantitative and qualitative feedback on using interim checklist version 2. The mean answer to the general question "On a scale from 1 to 10: how easy or difficult did you find it to use the checklist?" (1=very easy, 10=very difficult) was 4.0 (SD 1.8), with a range from 2 to 9. We asked participants to check the items they found difficult to answer, and four items were marked by either 8, 9, or 11 participants (40–55%, see Table 4). Explanations to why these items were difficult were also given (see Table 4). All this information was taken into consideration in our discussion on the final selection of items for the checklist. As evidenced in Table 4, none of the four items that were considered difficult by nearly half or more than half of the lay users made it to the final checklist. The reasons for the difficulties in applying certain items were manifold. For example, regarding item 23 "The health information states whether and how readers were involved in creating the information," over half of the lay users stated that it was either difficult to find that information on the webpage or the concept of reader involvement in creating health information was unclear to them. Furthermore, no one rated this item as especially important.

Conversely, four items were also rated as especially important by 9, 11, or 14 participants (45–70%, see Table 4), and three of these made it to the final checklist. Multiple users had trouble understanding item 14 "The health information indicates how well a fact is scientifically proven," and seven also marked it as a difficult item. However, more than half also rated it as especially important, and this item made it to the final checklist.

For item 19 "It is clear that the health information is up to date," one-third of participants demanded a clear timeframe for their judgment of up-to-dateness. Since the timeframe depends on the health information topic and is highly variable, this item was not included in the final checklist, but a more neutral item on currency was included ("It is clear when the health information was created or updated.").

Predictive validity with a test set and creating the final checklist

Out of 100 health information webpages in the test set, 37 were factually correct, that is, with no or little difference between the *claimed* and *factual* strength of evidence; 63 webpages were incorrect. The regression coefficients of the model for predicting the correctness of the health information webpages are shown in Table 4, along with the estimated error and 96% confidence interval. The higher the values, the better the respective item's predictive validity in predicting the trustworthiness of the online health information.

Table 4 also shows the agreement among lay users and among experts, respectively. Based on the initial results, we tested a few item combinations until we arrived at the final checklist (see Textbox 1).

As stated above, all quantitative and qualitative results were integrated into our final decision.

Some items with the highest regression coefficient for predicting a correct health information webpage were not chosen because of other reasons. For example, item 23 "The health information states whether and how readers were involved in creating the information" with the highest predictability values was only fulfilled in 8 out of 100 health information webpages. Therefore, the practical value of this item's predictive validity is low. Furthermore, more than half of the lay users perceived this item as difficult to answer, and the agreement among lay users in answering this item was very low (see Table 4). The situation was similar with item 22 "It is described how the information was created. For example – which studies were considered and why – and which were not." While this had a good predictive value, 8 out of 20 lay users perceived it to be difficult to answer; therefore, we decided not to include it in the final checklist, as only items that are easily answered and clear to laypeople can also be useful in the final checklist.

After several discussion rounds among the whole research team, we arrived at the final iWISE checklist with 7 items (see Textbox 1).

According to our statistical model, the probability that the health information is correct if all of these items were marked with "yes" is 99.9%. The probability that the health information is correct if all of the items are marked with "no" is 20.4%. However, this requires that the items are marked correctly by the user, which in practice may not be the case.

Textbox 1: Seven items of the final iWISE checklist

1. The health information does not contain advertising.
2. I feel the information is presented in a balanced way (the health information describes, for example, advantages and disadvantages, different treatment options, ...).
3. Technical terms are used sparingly, and their meaning is explained.
4. The health information is from an independent institution that presumably does not make any money from our health (e.g., no suppliers of medicines or food supplements, ...).
5. The health information states in detail which sources are behind the facts mentioned (bibliography, links to studies, ...).
6. The health information indicates how well the facts claimed are scientifically supported.
7. It is clear when the health information was created or updated.

We published the checklist on our webpage [39] and also provided short additional explanations for each item to further guide users in applying the checklist items. Furthermore, we created a video explaining all seven items to aid users in the checklist's application.

Discussion

Principal results and comparison with prior work

The iWISE checklist is a short, easy to understand, and easy to apply checklist for lay users to judge the trustworthiness of health information webpages. To our best knowledge, it is the first checklist that can be applied by laypersons in everyday situations without the need for prior training.

Numerous tools and checklists are available to assess online health information [13; 40-42]; however, these tools were either prepared for research purpose use or use by health care professionals, have not been developed by considering the lay user perspective, contain a large number of items, or have a sophisticated scoring system that makes it impossible for consumers to actually use the tool [12; 20; 43; 44].

Others are tools to judge written consumer information, such as the well-known DISCERN instrument [45; 46] or the Ensuring Quality Information for Patients (EQIP) tool [47; 48], neither of which can be directly transferred to the use for online health information.

In general, the wide range of tools for rating the quality of health-related websites can be broadly distinguished into different approaches [49]: 1) quality criteria for health information providers or developers that contain a list of recommendations for website development and content (e.g., [50-54]), 2) quality labels or logos that are displayed on screen and represent a provider's commitment to implement or adhere to a set of quality criteria (e.g., the widely used HON code [55], which was permanently discontinued in December 2022), and 3) user guidance systems for lay users or patients that enable them to check whether a site and its contents comply with certain standards by accessing a series of questions (e.g., [12; 13; 20; 43; 44]).

In the development of the iWISE checklist, we searched the literature comprehensively and accounted for all of those approaches as well as the literature on indicators, quality criteria, or the concept of online health information trustworthiness. We involved experts on health information and patient communication as well as laypersons with a low general education level and no formal health-related education. This was especially important in the development, as other quality evaluation checklists are mostly based on experts' views [56] and may not effectively serve lay users' needs. For example, the recently developed MAPPinfo checklist exclusively comprises items that are justified by either ethics or research evidence. However, it may therefore lack important criteria that are not yet evidence-based [12]. Furthermore, MAPPinfo user tests were done with health and nursing science students who found it easy to understand and use. In contrast, the iWISE checklist is made for laypersons without any prior medical or specialized health-related knowledge, and we involved lay users at several time points during the development process. We specifically eliminated all items that necessitated expertise or prior knowledge, so that the iWISE checklist can be applied only by reading the health-related webpage that one wants to appraise. Thereby we have ensured that the checklist items are both easy to understand and easy to apply for untrained laypersons. For everyday internet searches, it is often inconvenient to use a separate checklist because this requires either printing it out or constantly switching to it on the same electronic screen. A checklist is most useful when it is short enough to remember all of its items. Psychological research has long ago shown that the human short-term memory can hold seven items (plus or minus two) [57]. With seven items, our checklist is therefore short enough to remember.

The validation test showed that the items in the iWISE checklist can help distinguish between evidence-based and non-evidence-based online health information. The operationalization of health information trustworthiness that we chose to base our checklist validation on is novel and unique. As a proxy for the trustworthiness, we used the factual correctness of a particular health question presented on a webpage. The factual correctness was fulfilled if the *factual* strength of evidence for a certain health question and the *claimed* strength of evidence on the webpage were the same or only differed slightly. For our validation, we used the GRADE system [34], which is widely used in the field of evidence-based medicine, to rate the strength of evidence. Only items that were reasonably associated with factual correctness were included in the final iWISE checklist. Furthermore, the factual correctness is among the most important criteria of online health information for lay users [56].

The seven indicators and underlying criteria that constitute the iWISE checklist align with those found by Sun et al. 2019 [56], showing that our checklist is easy to relate to users/consumers and is congruent with their evaluation of quality. Sun et al. 2019 [56] summarized the relevant criteria and indicators that are important for the evaluation of online health information from the consumer’s perspective. They concluded that online health information quality could be reasonably measured by a small set of core dimensions that consumers deemed important. The three criteria that constitute the core dimensions of online health information quality as perceived by consumers were trustworthiness (whether a source or information can be trusted), expertise (whether a source or author has a sufficient level of subject-related knowledge), and objectivity (whether a source or information presents facts that are not influenced by personal feelings or commercial interests).

As pointed out by Hanif et al. [58], rating tools should be available to consumers, require a limited number of elements to be assessed, be assessable in all elements, be readable, and be able to gauge the readability and consistency of the information provided from a patient’s view point. These are all points that were taken into account in the iWISE checklist.

Limitations

Our checklist has some limitations. With 100 webpage articles, our validation test set of health information articles was relatively small. We tested our checklist only with a limited set of 10 popular health questions—and with 10 articles for each of these questions, respectively. Our test set might therefore be biased and not be representative of the broad mass of health information that a typical user encounters when searching the web. A different, more varied, and larger test set might therefore have resulted in a slightly different combination of checklist items. However, iWISE is comparable with other checklists that have either used a smaller validation test set or that deal with one health topic or both, for example, the MAPPinfo checklist was tested on 57 webpages [12], the QUality Evaluation Scoring Tool (QUEST) on 45 online articles on Alzheimer’s disease treatment and prevention [43], and Dobbins et al. used 120 web resources and another 107 in a second validation round on healthy aging [44].

The predictive validity results, one of the factors we considered in selecting the final checklist items, are based on the ratings from the application test with the research team. Laypersons' ratings were not integrated in this validation step. However, the perspectives of laypersons were considered in other ways, and we specifically ensured that all items were well understood by our lay user group.

Our checklist has only been validated for information on treatments or health decisions. It has not been tested for articles lacking interventional information. The checklist might therefore not be valid for judging the accuracy of information on symptoms for the diagnosis of a disease or on the health risks of non-interventional hazard exposure. Also, our checklist is not tested with social media postings or health information videos and cannot be used for evaluating medication leaflets.

Conclusions

The iWISE checklist is a user-friendly, easy-to-understand checklist for evaluating the trustworthiness of health information websites. It has been carefully developed and validated. The perspective of lay users has been especially integrated into the final checklist. With only seven items, the final checklist is easy enough to remember and could significantly improve critical health literacy. The iWISE checklist is a tool that can empower individuals to make more informed decisions about their health while potentially reducing the spread of misinformation.

In future studies, the reliability of the iWISE checklist should be tested, and it should be tested for use on social media posts or health information videos.

Acknowledgments

We thank Evelyn Auer for excellent help in project administration and Isolde Sommer for her contributions to the data extraction and expert Delphi study. We further thank Genevieve Iseult Eldredge for proofreading. During the manuscript preparation, the authors used DeepL, Claude 3.5 Sonnet, and ChatGPT for help with text improvements and the translations from German to English. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

This study was supported by internal funding from the University of Continuing Education Krems, Austria.

Author contributions

BK and UG conceived the idea and led the grant application for funding. UG led the project planning and conduct. UG, BK, CK, IK, BL, EK, and FKA wrote the study protocol. IK and BK performed the electronic database searches; IK, BK, FKA, CK, BL, EK, and UG performed the supplementary searches and screened the literature. BK, CK, EK, FKA, IK, and UG did the data extractions, CK, EK, IK, and UG performed the categorization and first condensation of the items. EK, FKA, BL, UG, BK, CK, IK, and IS did the item rating for the second condensation step. CK planned and carried out the Delphi study. UG, CK, and EK carried out and analyzed the cognitive interviews. UG led the application test with the lay users; UG, DL, and IK analyzed the data. BK led the operationalization of health information trustworthiness and compiled the webpages test set. BK and IM assessed the strength of evidence for the test set webpages. RE planned and conducted the statistical analyses for the predictive validity. All authors engaged in discussions and decisions on the final checklist. All authors drafted the article, critically revised it for important intellectual content, and approved the final version for publication. UG and BK act as guarantors. The corresponding author attests that all listed authors meet the authorship criteria and that no others meeting the criteria have been omitted.

Conflicts of interest

None declared.

Abbreviations

BASE: Bielefeld Academic Search Engine
EQIP: Ensuring Quality Information for Patients
GRADE: Grading of Recommendations, Assessment, Development and Evaluation
HLS-EU: Health Literacy Survey in Europe
iWISE: Info without Side Effects
MAPPinfo: Mapping the Quality of Health Information
OSF: Open Science Framework
QUEST: QUality Evaluation Scoring Tool
SRQR: Standards for Reporting Qualitative Research

References

1. Eurostat. Individuals using the internet for seeking health-related information 2024: https://ec.europa.eu/eurostat/databrowser/view/tin00101/default/table?lang=en (Accessed: 10 Sep 2024).

2. Griebler R, Straßmayr C, Mikšová D, Link T, Nowak P, Arbeitsgruppe Gesundheitskompetenz-Messung der ÖPGK. Gesundheitskompetenz in Österreich: Ergebnisse der österreichischen Gesundheitskompetenz HLS19-AT. Wien: Bundesministerium für Soziales, Gesundheit, Pflege und Konsumentenschutz; 2021.

3. Daraz L, Morrow AS, Ponce OJ, Beuschel B, Farah MH, Katabi A, et al. Can Patients Trust Online Health Information? A Meta-narrative Systematic Review Addressing the Quality of Health Information on the Internet. Journal of General Internal Medicine. 2019;34(9):1884-91 DOI: https://doi.org/10.1007/s11606-019-05109-0.

4. The HSL19 Consortium of the WHO Action Network M-POHL. International report on the methodology, results, and recommendations of the European health literacy population survey 2019-2021 (HLS19) of M-POHL 2021: https://m-pohl.net/sites/m-pohl.net/files/inline-files/HLS19%20International%20Report.pdf (Accessed: 31 Aug 2023).

5. Marstedt G. Das Internet: Auch Ihr Ratgeber für Gesundheitsfragen? Bevölkerungsumfrage zur Suche von Gesundheitsinformationen im Internet und zur Reaktion der Ärzte 2018: https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/VV_Studie_Das-Internet-auch-Ihr-Ratgeber_Befragung.pdf (Accessed: 31 Aug 2023).

6. Sørensen K, Pelikan JM, Röthlin F, Ganahl K, Slonska Z, Doyle G, et al. Health literacy in Europe: comparative results of the European health literacy survey (HLS-EU). European journal of public health. 2015;25(6):1053-8 DOI:

https://doi.org/10.1093/eurpub/ckv043.

7. Kickbusch I, Pelikan JM, Apfel F, Tsouros AD. Health literacy. The solid facts. 2013: https://iris.who.int/bitstream/handle/10665/128703/e96854.pdf (Accessed: 31 Aug 2023).

8. Nutbeam D. The evolving concept of health literacy. Social science & medicine. 2008;67(12):2072-8. PMID: 18952344.

9. Diviani N, van den Putte B, Giani S, van Weert JC. Low Health Literacy and Evaluation of Online Health Information: A Systematic Review of the Literature. J Med Internet Res. 2015;17(5):e112 DOI: https://doi.org/10.2196/jmir.4018

10. United Nations. Goals 3: Ensure healthy lives and promote well-being for all at all ages. Overview 2024: https://sdgs.un.org/goals/goal3 (Accessed: 1 Dec 2024).

11. Sykes S, Wills J, Rowlands G, Popple K. Understanding critical health literacy: a concept analysis. BMC public health. 2013;13:150 DOI: https://doi.org/10.1186/1471-2458-13-150.

12. Kasper J, Lühnen J, Hinneburg J, Siebenhofer A, Posch N, Berger-Höger B, et al. MAPPinfo - mapping quality of health information: Validation study of an assessment instrument. PLOS ONE. 2023;18(10):e0290027 DOI: https://doi.org/10.1371/journal.pone.0290027.

13. Kasper J, Lühnen J, Hinneburg J, Siebenhofer A, Posch N, Berger-Höger B, et al. MAPPinfo, mapping quality of health information: study protocol for a validation study of an assessment instrument. BMJ Open. 2020;10(11):e040572 DOI: https://doi.org/10.1136/bmjopen-2020-040572.

14. Griebler U, Kerschner B, Kien C, Klerings I, Lutz B, Krczal E, et al. Development of a quality criteria catalogue for evaluating online health information and the corresponding training interventions for lay people – Study protocol. University for Continuing Education Krems. 2022: https://osf.io/bfwqh.

15. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for Reporting Qualitative Research: A Synthesis of Recommendations. Academic Medicine. 2014;89(9):1245-51 DOI: https://doi.org/10.1097/acm.0000000000000388.

16. American Psychological Association. Mixed Methods Article Reporting Standards (MMARS) 2020: https://apastyle.apa.org/jars/mixed-methods (Accessed: 27 Nov 2024).

17. Viviani M, Pasi G. Credibility in social media: opinions, news, and health information—a survey. WIREs Data Mining and Knowledge Discovery. 2017;7(5):e1209 DOI: https://doi.org/10.1002/widm.1209.

18. Cambridge Dictionary. Meaning of "layperson" in English 2023: https://dictionary.cambridge.org/dictionary/english/layperson (Accessed: 28 March

2022).

19. Cooper C, Booth A, Husk K, Lovell R, Frost J, Schauberger U, et al. A Tailored Approach: A model for literature searching in complex systematic reviews. Journal of Information Science. 2024;50(4):1030-62 DOI: https://doi.org/10.1177/0165551522111445.

20. Provost M, Koompalum D, Dong D, Martin BC. The initial development of the WebMedQual scale: Domain assessment of the construct of quality of health web sites. International Journal of Medical Informatics. 2006;75(1):42-57 DOI: https://doi.org/10.1016/j.ijmedinf.2005.07.034.

21. Powell C. The Delphi technique: myths and realities. J Adv Nurs. 2003;41(4):376-82 DOI: https://doi.org/10.1046/j.1365-2648.2003.02537.x.

22. Barrett D, Heale R. What are Delphi studies? Evid Based Nurs. 2020;23(3):68-9 DOI: https://doi.org/10.1136/ebnurs-2020-103303.

23. Jobe JB. Cognitive psychology and self-reports: Models and methods. Quality of Life Research. 2003;12(3):219-27 DOI: https://doi.org/10.1023/A:1023279029852.

24. Willis GB. Analysis of the cognitive interview in questionnaire design. New York: Oxford University Press; 2015. ISBN 978-0-19-995775-0

25. Scott K, Ummer O, LeFevre AE. The devil is in the detail: reflections on the value and application of cognitive interviewing to strengthen quantitative surveys in global health. Health Policy and Planning. 2021;36(6):982-95 DOI: https://doi.org/10.1093/heapol/czab048.

26. Beatty PC, Willis GB. Research Synthesis: The Practice of Cognitive Interviewing. Public Opinion Quarterly. 2007;71(2):287-311 DOI: https://doi.org/10.1093/poq/nfm006.

27. Buers C, Triemstra M, Bloemendal E, Zwijnenberg NC, Hendriks M, Delnoij DMJ. The value of cognitive interviewing for optimizing a patient experience survey. International Journal of Social Research Methodology. 2014;17(4):325-40 DOI: https://doi.org/10.1080/13645579.2012.750830.

28. Willis G. Pretesting of Health Survey Questionnaires: Cognitive Interviewing, Usability Testing, and Behavior Coding. In: Health Survey Methods; Johnson TP. ed. 2014. p. 217-42.

29. Patton MQ. Qualitative research & evaluation methods: Integrating theory and practice: Sage publications; 2014. ISBN 1483301451

30. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Medical Research Methodology. 2013;13(1):117 DOI: https://doi.org/10.1186/1471-2288-13-117.

31. Cochrane Austria. Department for Evidence-Based Medicine and Evaluation. Medizin-Transparent 2024: https://medizin-transparent.at/ (Accessed: 10 Sep 2024).

32. International Fact-Checking Network (IFCN). Medizin transparent - Universität für Weiterbildung Krems (Donau-Universität Krems) 2024: https://ifcncodeofprinciples.poynter.org/profile/medizin-transparent-universitat-fur-weiterbildung-krems-donau-universitat-krems (Accessed: 10 Sep 2024).

33. Cochrane Austria. Department for Evidence-Based Medicine and Evaluation. Methodenpapier Medizin-transparent. Version 1.0 vom 5. April 2018 2018: https://medizin-transparent.at/wp-content/uploads/2022/12/Methodenpapier-Medizin-transparent_v1.0.pdf (Accessed: 10 Sep 2024).

34. Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011;64(4):401-6 DOI: https://doi.org/10.1016/j.jclinepi.2010.07.015.

35. Santesso N, Glenton C, Dahm P, Garner P, Akl EA, Alper B, et al. GRADE guidelines 26: informative statements to communicate the findings of systematic reviews of interventions. J Clin Epidemiol. 2020;119:126-35 DOI: https://doi.org/10.1016/j.jclinepi.2019.10.014.

36. Kerschner B, Wipplinger J, Klerings I, Gartlehner G. [How evidence-based are print- and online mass media in Austria? A quantitative analysis] Wie evidenzbasiert berichten Print- und Online-Medien in Österreich? Eine quantitative Analyse. Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen. 2015;109(4):341-9 DOI: https://doi.org/10.1016/j.zefq.2015.05.014.

37. Bürkner P-C. Bayesian Item Response Modeling in R with brms and Stan. Journal of Statistical Software. 2021;100(5):1-54 DOI: https://doi.org/10.18637/jss.v100.i05.

38. Wickham H, Averick M, Bryan J, Chang W, McGowan LDA, François R, et al. Welcome to the Tidyverse. Journal of open source software. 2019;4(43):1686 DOI: https://doi.org/10.21105/joss.01686.

39. Infos Ohne Nebenwirkung. Checkliste für verlässliche Gesundheitsinfos 2024: https://www.infos-ohne-nebenwirkung.at/checkliste/ (Accessed: 27 Nov 2024).

40. Bernstam EV, Shelton DM, Walji M, Meric-Bernstam F. Instruments to assess the quality of health information on the World Wide Web: what can our patients actually use? Int J Med Inform. 2005;74(1):13-9 DOI: https://doi.org/10.1016/j.ijmedinf.2004.10.001.

41. Gagliardi A, Jadad AR. Examination of instruments used to rate quality of health information on the internet: chronicle of a voyage with an unclear destination. BMJ. 2002;324(7337):569-73 DOI: https://doi.org/10.1136/bmj.324.7337.569.

42. Song S, Zhang Y, Yu B. Interventions to support consumer evaluation of online health information credibility: A scoping review. International Journal of Medical Informatics. 2021;145:104321 DOI: https://doi.org/10.1016/j.ijmedinf.2020.104321.

43. Robillard JM, Jun JH, Lai JA, Feng TL. The QUEST for quality online health information:

validation of a short quantitative tool. BMC Med Inform Decis Mak. 2018;18(1):87 DOI: https://doi.org/10.1186/s12911-018-0668-9.

44. Dobbins M, Watson S, Read K, Graham K, Yousefi Nooraie R, Levinson AJ. A Tool That Assesses the Evidence, Transparency, and Usability of Online Health Information: Development and Reliability Assessment. JMIR Aging. 2018;1(1):e3 DOI: https://doi.org/10.2196/aging.9216.

45. Charnock D. The DISCERN Handbook. Abingdon, Great Britain: Radcliff Medical Press, University of Oxford and The British Library; 1998. ISBN 1 85775 310 0

46. 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(2):105-11. PMID: 1756830.

47. Moult B, Franck LS, Brady H. Ensuring Quality Information for Patients: development and preliminary validation of a new instrument to improve the quality of written health care information. Health Expectations. 2004;7(2):165-75 DOI: https://doi.org/10.1111/j.1369-7625.2004.00273.x.

48. Charvet-Berard AI, Chopard P, Perneger TV. Measuring quality of patient information documents with an expanded EQIP scale. Patient Education and Counseling. 2008;70(3):407-11 DOI: https://doi.org/10.1016/j.pec.2007.11.018.

49. Wilson P. How to find the good and avoid the bad or ugly: a short guide to tools for rating quality of health information on the internet. Bmj. 2002;324(7337):598-602 DOI: https://doi.org/10.1136/bmj.324.7337.598.

50. 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. American Medical Association. Jama. 2000;283(12):1600-6 DOI: https://doi.org/10.1001/jama.283.12.1600.

51. Minervation. The LIDA Instrument 2008: https://www.minervation.com/wp-content/uploads/2011/04/Minervation-LIDA-instrument-v1-2.pdf (Accessed: 6 Dec 2021).

52. Silberg WM, Lundberg GD, Musacchio RA. Assessing, controlling, and assuring the quality of medical information on the Internet: Caveant lector et viewor--Let the reader and viewer beware. JAMA. 1997;277(15):1244-5. PMID: 9103351.

53. Rippen H, Risk A. e-Health Code of Ethics (May 24). J Med Internet Res. 2000;2(2):e9 DOI: https://doi.org/10.2196/jmir.2.2.e9.

54. Leitlinie Evidenzbasierte Gesundheitsinformation. Leitlinie Evidenzbasierte Gesundheitsinformation 2024: https://www.leitlinie-gesundheitsinformation.de/ (Accessed: 26 Nov 2024).

55. Boyer C, Baujard V, Geissbuhler A. Evolution of health web certification through the

HONcode experience. Stud Health Technol Inform. 2011;169:53-7. PMID: 21893713.

56. Sun Y, Zhang Y, Gwizdka J, Trace CB. Consumer Evaluation of the Quality of Online Health Information: Systematic Literature Review of Relevant Criteria and Indicators. J Med Internet Res. 2019;21(5):e12522 DOI: https://doi.org/10.2196/12522.

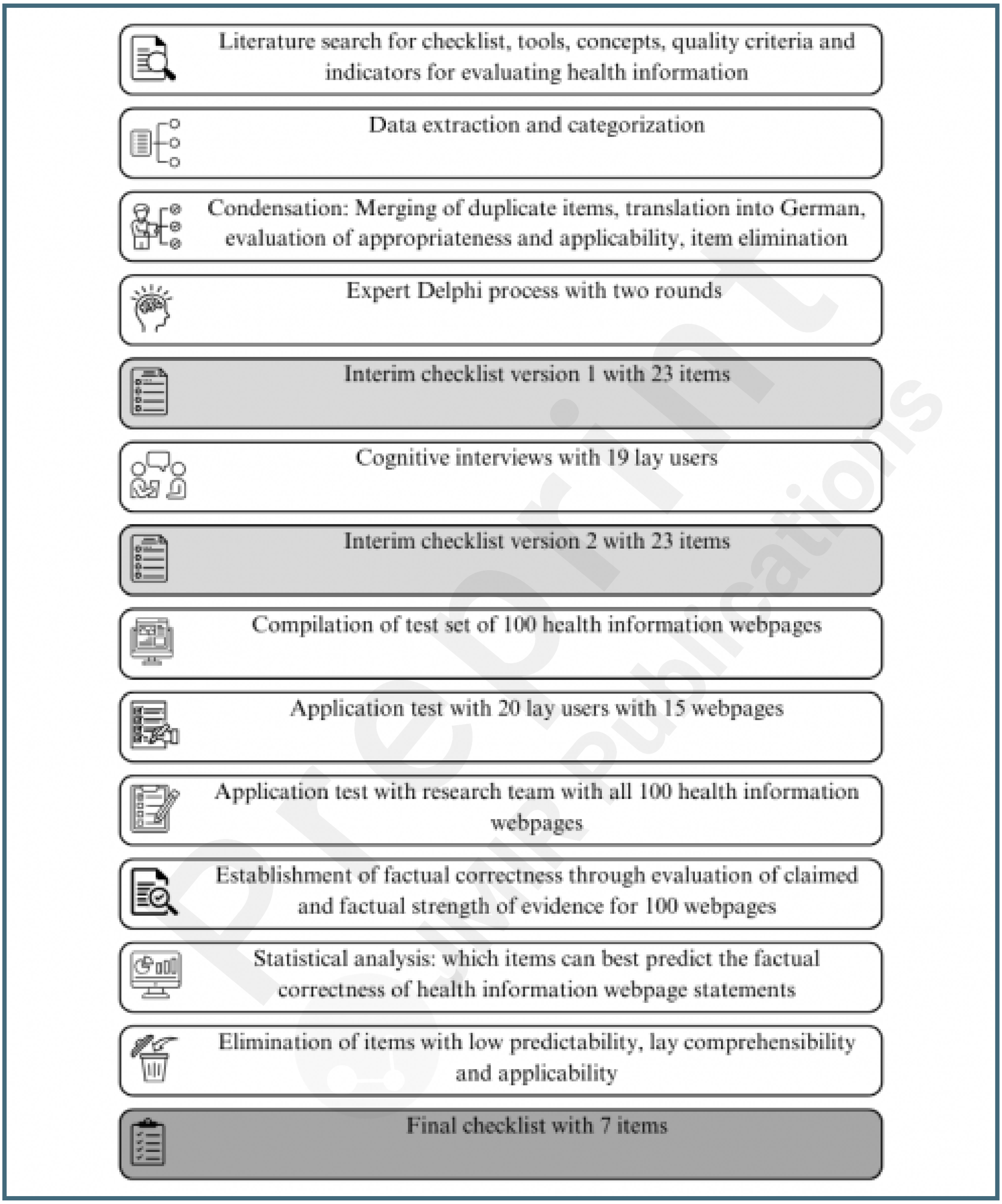
57. Miller GA. The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological review. 1956;63(2):81. PMID: 13310704.

58. Hanif F, Read JC, Goodacre JA, Chaudhry A, Gibbs P. The role of quality tools in assessing reliability of the internet for health information. Inform Health Soc Care. 2009;34(4):231-43 DOI: https://doi.org/10.3109/17538150903359030.

Supplementary Files

Figures

Overview of the iWISE checklist development process.



Literature search for checklist, tools, concepts, quality criteria and indicators for evaluating health information

Data extraction and categorization

Condensation: Merging of duplicate items, translation into German, evaluation of appropriateness and applicability, item elimination

Expert Delphi process with two rounds

Interim checklist version 1 with 23 items

Cognitive interviews with 19 lay users

Interim checklist version 2 with 23 items

Compilation of test set of 100 health information webpages

Application test with 20 lay users with 15 webpages

Application test with research team with all 100 health information webpages

Establishment of factual correctness through evaluation of claimed and factual strength of evidence for 100 webpages

Statistical analysis: which items can best predict the factual correctness of health information webpage statements

Elimination of items with low predictability, lay comprehensibility and applicability

Final checklist with 7 items

Multimedia Appendixes

Supplementary tables A1 to A4.
URL: http://asset.jmir.pub/assets/7dabe95766938c811b8e8ae53aa1e5ac.docx