PLOS Digital Health

RESEARCH ARTICLE

The trust in AI-generated health advice (TAIGHA) scale and short version (TAIGHA-S): Development and validation study

Marvin Kopka[1,2*], **Azeem Majeed**[3], **Gabriella Spinelli**[4], **Austen El-Osta**[2,5‡], **Markus Feufel**[1‡]

1 Division of Ergonomics, Department of Psychology and Ergonomics (IPA), Technische Universität Berlin, Berlin, Germany, **2** Self-Care Academic Research Unit (SCARU), School of Public Health, Imperial College London, London, United Kingdom, **3** Department of Public Health and Primary Care, Imperial College London, London, United Kingdom, **4** College of Engineering, Design and Physical Sciences, Brunel University of London, Uxbridge, United Kingdom, **5** School of Life Course and Population Sciences, King's College London, London, United Kingdom

‡ These authors share the last authorship.
* m.kopka@imperial.ac.uk

Abstract

Artificial Intelligence (AI) tools such as large language models (LLMs) are increasingly used by the public to obtain health information and support health-related decisions. Because such use involves advice-taking, following or rejecting AI-generated advice can have direct clinical and safety implications and consequences for the healthcare system. Although trust plays an important role in the adoption of health-related AI advice, existing instruments only assess trust in generic technology and perceived trustworthiness. There are currently no validated instruments that specifically measure users' state trust in AI-generated health advice, and an alternative for self-developed one-item scales is missing.This study aimed to develop and validate the Trust in AI-Generated Health Advice (TAIGHA) scale and its four-item short form (TAIGHA-S) as theory-based questionnaires for measuring trust and distrust in AI-generated health advice. We used a generative AI approach to generate new use-case specific candidate items based on existing theory that each comprised cognitive and affective components. After automated validation, we conducted manual validation in three steps: (i) content validation with ten domain experts, (ii) face validation with 30 lay participants, and (iii) psychometric validation with 385 UK participants receiving AI-generated health advice for symptom assessment. After automated item reduction, 28 items were retained and further reduced to 10 based on expert ratings. The final TAIGHA scale showed excellent content validity (S-CVI/Ave = 0.99) and face validity (S-FVI/Ave = 0.99). CFA confirmed a two-factor model with excellent fit (CFI = 0.98, TLI = 0.98, SRMR = 0.03). Internal consistency was high ($\alpha = 0.94$, $\omega = 0.94$ for trust; $\alpha = 0.93$, $\omega = 0.93$ for distrust). The short form correlated highly with the full scale ($r = 0.96$) and showed high reliability ($\alpha = 0.88$, $\omega = 0.89$ for

OPEN ACCESS

Citation: Kopka M, Majeed A, Spinelli G, El-Osta A, Feufel M (2026) The trust in AI-generated health advice (TAIGHA) scale and short version (TAIGHA-S): Development and validation study. PLOS Digit Health 5(7): e0001488. https://doi.org/10.1371/journal.pdig.0001488

Editor: Anna Khalemsky, Hadassah Academic College, ISRAEL

Received: December 17, 2025

Accepted: May 22, 2026

Published: July 2, 2026

Data availability statement: The data is openly accessible via Zenodo: https://doi.org/10.5281/zenodo.19355964.

Funding: The author(s) received no specific funding for this work.

Competing interests: AEO & GS are co-investigators of the RADIANT Centre for Excellence in Regulatory Science and Innovation (CERSI), funded by the Medical Research Council (MRC), Innovate UK, the Medicines and Healthcare products Regulatory Agency (MHRA) and the Office for Life Sciences (OLS).

Abbreviations: AI, Artificial Intelligence; CFA, Confirmatory Factor Analysis; CFI, Comparative Fit Index; CVI, Content Validity Index; CVI-I, Item-Level Content Validity Index; DST, Decision Support Tool; FVI, Face Validity Index; FVI-I, Item-Level Face Validity Index; GFI: Goodness of Fit Index; IQR, Interquartile Range; ITC, Item-Total Correlation; LLM, Large Language Model; M (SD), Mean (Standard Deviation); NASA-TLX, NASA Task Load Index; NFI, Normed Fit Index; NMI, Normalized Mutual Information; RMSEA, Root Mean Square Error of Approximation; SRMR, Standardized Root Mean Squared Residual; S-CVI/Ave, Scale-Level CVI, Average; S-CVI/UA, Scale-Level CVI, Universal Agreement; S-FVI/Ave, Scale-Level FVI, Average; S-FVI/UA, Scale-Level FVI, Universal Agreement; TAIGHA, Trust in AI-Generated Health Advice (Scale); TAIGHA-S, TAIGHA Short Form; TLI, Tucker-Lewis Index; UK, United Kingdom.

trust; $\alpha = 0.84$, $\omega = 0.85$ for distrust). TAIGHA and TAIGHA-S are validated instruments for assessing users' state trust and distrust in AI-generated health advice, with excellent psychometric properties and stronger associations with reliance than existing general trust scales.

Author summary

Artificial intelligence tools are increasingly used by the public to obtain health information and guidance, but users may follow or reject the advice they receive, which can affect their health and the healthcare system. Although trust is very important, most existing questionnaires measure trust in technology in general, and not specifically for health advice produced by artificial intelligence. In this study, we developed and validated a new questionnaire that is specific to this context: the Trust in AI-Generated Health Advice (TAIGHA) scale, along with a short four-item version for situations where time is limited. We first generated candidate questions based on established theory and then refined them through a stepwise evaluation with experts and members of the public. Finally, we tested the questionnaire with 385 adults in the United Kingdom who received artificial intelligence-generated advice in a symptom-assessment scenario. The final scale showed good properties to be used as a tool for measuring trust in AI-generated health advice. By providing a practical way to measure both trust and distrust separately, our scale can support evaluations of health-focused artificial intelligence and help researchers and developers understand when and why people rely on such advice.

Introduction

Given the growing popularity, availability and performance of generative Artificial Intelligence (AI) tools such as Large Language Models (LLMs), the public are increasingly using these technologies to obtain health information and guidance for a variety of health-related tasks and decisions [1]. This growing reliance on AI-generated information is important in health-related contexts, where following or rejecting an AI tool's advice can have personal, clinical and safety consequences, as well as broader impacts on healthcare systems [2–4]. The recent case of a ChatGPT user who was hospitalised for bromism after following advice on how to reduce salt intake demonstrated these risks [5]. Similarly, LLMs may also inadvertently spread misinformation when generating inaccurate or fabricated content [6,7]. Whereas such incidents exemplify potential dangers, the same technology also promises to make healthcare more efficient.

Emerging empirical evidence suggests that these risks are amplified by users' high levels of trust in AI-generated medical advice. A recent MIT study [8] found that patients often trust medical recommendations produced by AI systems more than those provided by human clinicians, even when the AI advice is demonstrably incorrect. Notably, participants were less likely to critically challenge AI-generated guidance and more inclined to follow it with confidence, raising concerns about overreliance and reduced scepticism in decision-making. This tendency is particularly problematic in health contexts, where misplaced trust may lead to harmful self-management behaviours, delayed clinical intervention, or inappropriate treatment decisions.

At a system level, for instance, LLMs may support patient empowerment, their decision-making, and health education in community settings [9–12]. For non-experts, determining the accuracy of information or advice provided by an AI decision support tool (DST) is often challenging, yet they must still decide whether to trust and then follow the AI-generated advice. Although trust in AI has been shown to be an important predictor of whether people will use AI tools and follow, or not, their advice [10,13], trust formation is context-dependent and may differ depending on personal and situational variables [14]. Measuring use-case specific trust is therefore important to understand factors associated with how users respond to AI-generated advice. Because no existing instrument specifically measures the trust that lay users place in AI-generated health advice, we sought to develop and validate the Trust in AI-Generated Health Advice (TAIGHA) scale.

Research on the measurement of trust in technology has its roots in organizational trust measurement. In one of the earliest articles, Mayer and colleagues defined trust as "*the willingness to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor*" [15]. According to this definition, trust comprises three dimensions: (i) competence, (ii) integrity, and (iii) benevolence [15]. An alternative conceptualisation agrees with the general definition but distinguishes between cognitive and affective trust [16]. Whereas cognitive trust refers to positive beliefs about the trustee's ability and reliability, affective trust refers to the emotional dimension and perceived care the trustee demonstrates toward the trustor [16].

Trust in technology is most commonly assessed using the Trust in Automated Systems Survey, which is the most frequently cited instrument in this field [17,18]. This scale builds on Mayer and colleagues' definition [15] and asks users to rate their perceptions of a technological system. This scale has several limitations: it tends to produce overly positive responses [19], comprises two subscales rather than a single construct as originally proposed [20], and lacks specificity due to its broad and context-independent wording [18]. Although other psychometrically validated instruments have since been developed, including the Human-Computer Trust Scale [21], the Semantic Differential Scale for AI Trust [22], and the Trust Scale for Explainable AI [23], these instruments share similar limitations. Most importantly, they are not sensitive to specific use cases as they measure general trust in technology rather than context-dependent trust.

As the reliance on AI-based DSTs continues to grow across diverse domains, the context in which they are used also impacts how trust should be measured. That is, a scale for evaluating trust in autonomous vehicles should include different questions than a scale on trust in an AI system that provides medical advice. In other words, the context of implementation is important to develop trust measurement scales with higher (use-case specific) applicability. Similarly, Schlicker et al. argue that most existing scales do not actually measure trust, but rather a general perceived trustworthiness [24]. Trustworthiness refers to properties of a system (e.g., the developer, the algorithms, its perceived accuracy), whereas trust refers to the attitude a person has towards this system. Most validated instruments, however, confound trustworthiness and trust and validated instruments for actual trust are missing, which is particularly relevant in health DST use cases, where users may perceive an AI tool as not trustworthy but still decide to follow its advice [8,24]. It is therefore worthwhile to consider how trust should be measured in AI systems deployed to provide healthcare advice to a lay audience. In such contexts, users typically have no visibility of the developer, the provenance of the data, or the type of algorithm employed, and their only relationship is with the technology itself. As a result, trust is not mediated through institutional or professional actors but is formed directly through interaction with the AI system. Thus, measurement approaches that capture behavioural reliance rather than perceived system qualities alone are needed.

In summary, although several instruments exist to measure trust in technology, most measure perceived trustworthiness with limited applicability to the context of AI-generated health advice. Trust in this domain forms against the background of a particularly high degree of uncertainty due to use-case specific factors related to personal safety risks, the varying accuracy and quality of medical evidence and the differing transparency of AI advice. Hence, due to the nature of the AI technology and the opacity of the underlying decision-making process, asking its users rather general questions related to whether "the system behaves in an underhanded manner" [17], may not be meaningfully answered in the context of AI health advice. To address this gap, we developed and validated the Trust in AI-Generated Health Advice (TAIGHA) scale.

To specifically measure state trust regarding a specific piece of health advice rather than perceived trustworthiness [24], the TAIGHA scale builds on McAllister's definition and conceptualises trust as consisting of both cognitive and affective components [16]. We did not build on the definition by Mayer et al., because it mostly contains attributes related to trustworthiness rather than trust [15]. The scale further builds on results from newer psychometric analyses of the original Trust in Automated Systems Survey, which identified two related dimensions: trust and distrust [20,25]. Accordingly, the TAIGHA scale provides a use-case-specific measure of state trust in a piece of AI-generated health advice (as opposed to broader trust in health AI tools) that includes two subscales, trust and distrust, each of which comprises cognitive and affective items. In addition to the main scale, we aimed to develop a short version, the TAIGHA-S, to measure trust and distrust in time-sensitive scenarios using as few items as possible. In applied research, trust is often measured as a secondary outcome using self-developed, single-item measures [26,27]. The TAIGHA-S is thus intended to offer researchers and practitioners a psychometrically validated alternative for situations in which they want to assess trust and distrust in AI-generated health advice and would otherwise rely on unvalidated questions.

The primary aim of this study was to develop and validate the TAIGHA scale, a psychometrically sound, context-specific instrument measuring trust and distrust in AI-generated health advice. Further to evaluating the factor structure, reliability and validity of the TAIGHA scale, we sought to develop and validate a short version (TAIGHA-S) for practical or large-scale use.

Methods

Ethics

This study received ethical approval from the Ethics Committee of the Department of Psychology & Ergonomics at Technische Universität Berlin (approval number 2708652). The study and procedures adhered to the Declaration of Helsinki and all participants provided written informed consent.

Instrument development

To develop the TAIGHA items, we used a generative AI-based method called AI-GENIE to develop an initial set of items and reduce this item set. In a manual validation, we then tested the items, reduced the item set further, and assessed validity and reliability. The full procedure is shown in Fig 1.

The AI-GENIE generative AI-based method [28], which uses an LLM to generate several candidate items and then obtains vector embeddings (numerical representations of text) for these items using an encoder, was used to develop the baseline version of the TAIGHA scale. More information on this approach can be found in [28]. In this process, the AI-GENIE represents the contextual meaning of each item numerically, which allows it to cluster items with similar meanings. For the vector embeddings, we used OpenAI's text-embedding-3-small. Next, the psychometric properties of the scale were assessed by AI-GENIE using a network-based method called exploratory graph analysis, which analyzes the vector embeddings to identify a factorial structure [29]. Subsequently, the normalized mutual information (NMI) value – which indicates the proportion of items classified into the same factor as in the input factorial structure [30] – was automatically calculated. At this stage, the NMI is expected to be relatively low, as no item selection has yet been applied. Additionally, the item pool may contain items that are phrased very similarly. To address this, unique variable selection was applied by AI-GENIE to remove items that were too similar in wording [31]. In the proceeding step, bootstrapped exploratory graph analysis was used by AI-GENIE to select only those items that were stable, i.e., items that consistently get grouped into the same factor. Here, item stability is calculated as a metric to determine how often an item gets

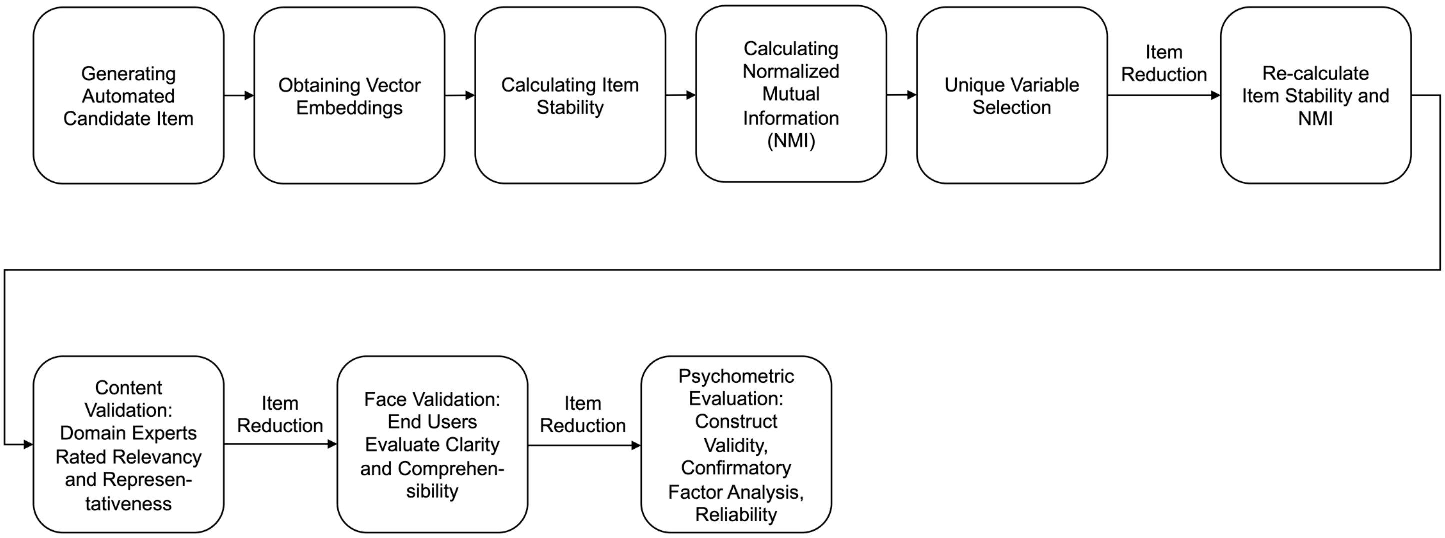

Fig 1. Instrument development and validation procedure.

https://doi.org/10.1371/journal.pdig.0001488.g001

grouped into the same category. After the item pool has been reduced using bootstrapped exploratory graph analysis, the reduced item pool was validated again using NMI and item stability by AI-GENIE.

We used GPT-4.1 within AI-GENIE to generate the items, because this model follows instructions consistently [32]. We applied the trust theory consisting of trust and distrust as two separate but related factors [20,25]. For the LLM input, trust was defined as the "willingness to follow health advice from an AI because you expect it to help you", and included cognitive items measuring the "expectation that following the AI's advice will lead to positive health outcomes", and affective items measuring "positive feelings such as comfort, reassurance, or peace of mind when considering or relying on the AI's advice itself". Distrust was defined as a "protective stance, avoiding this piece of health advice because you expect it to harm you", and included cognitive items measuring the "expectation that following the AI's advice will lead to negative health outcomes", and affective items measuring "negative feelings such as anxiety, unease, or worry when considering or relying on the AI's advice". The settings used in item generation can be found in Textbox A in S3 Appendix.

Although the automated validation procedure with the AI-GENIE may yield results that are highly similar to empirical validations [28] and on average produces even better items than those written by humans [33], empirical validation is still essential to generate valid scales. For example, the AI-GENIE approach can establish a factorial structure but generates items based solely on the input theory. Hence, the construct and content validity of the resulting items must be verified empirically. To include only items with high content validity (as judged by experts) and acceptable psychometric properties, we therefore conducted a traditional psychometric evaluation in the next step.

Validation procedure

We followed the validation procedure by Yusoff et al. [34]: first, we conducted content validation by surveying domain experts to determine whether each item was relevant and representative of the trust and distrust constructs. Second, we performed face validation by asking intended end users (i.e., general survey participants without psychometric or domain expertise) to evaluate whether each item was clear and comprehensible. Finally, after selecting the final item set, we conducted a psychometric evaluation where participants received health advice from an AI, made a health-related decision, and completed the TAIGHA scale along with related and unrelated questionnaires. Using these responses, we determined construct validity through correlations with related and unrelated measures, conducted a confirmatory factor analysis

(CFA) to assess whether the factorial structure fits with our theoretical structure, and also assessed reliability using Cronbach's Alpha and McDonald's Omega.

In the first step, ten domain experts completed an online questionnaire hosted on SoSciSurvey. They received an introduction to the theory the TAIGHA scale is based on, reviewed the items and rated each item's relevance and representativeness on a 4-point Likert scale.

In the second step, 30 non-experts completed a similar online questionnaire. They received a brief explanation of the study's purpose, a short introduction to the theory and an example of AI-generated health advice. They were then asked to rate each item's clarity and comprehensibility on a 4-point Likert scale.

In the third step, we used a symptom-assessment scenario previously used in experimental and observational studies [9]. We specifically selected this scenario because it represents a real-world use case of obtaining health advice from an AI-based DST, which is common among medical lay users [35–38]. We used a set of validated case scenarios that describe real-world patient cases in which medical lay users sought online health information to assist them in deciding whether and where to seek care [39,40]. After providing informed consent and demographic information, participants received one randomly selected scenario from the full set of 27 scenarios and were asked to indicate whether they would (i) seek emergency care for the described symptoms (i.e., call 999 or visit the emergency room), (ii) seek non-emergency care (i.e., trying to see a doctor within the next days), or (iii) engage in self-care (i.e., let the health issue get better on its own and review the situation in a few days again). They then received AI-generated health advice on what to do in this scenario (generated using GPT o3) and were asked to choose again among the three options. Finally, participants were asked to complete the TAIGHA scale, as well as several additional questionnaires measuring related and unrelated constructs.

Because participants complete the TAIGHA scale after viewing one advice instance, the present validation concerns a state measure related to a specific piece of advice rather than a trait or a longitudinal measure of trust.

Participants

Experts for the content validation were recruited via snowball sampling from the researchers' professional network. Following Yusoff et al. [41], we aimed to include ten experts, who participated between 27 May 2025 and 11 June 2025.

Participants for the face validation were recruited via Prolific [42] using a random sample of English-speaking users from the UK. In line with Yusoff et al. [43], we aimed to include 30 participants. Data for the face validation were collected on 11 June 2025.

Participants for the psychometric validation were also recruited via Prolific using a random sample of English-speaking users from the UK. Since there are no clear guidelines for sample size in CFA, but at least 300 participants are generally recommended [44], the study oversampled by 100 participants to account for the exclusion of participants that did not answer two embedded attention checks correctly. The recruitment target was a total of 400 participants by 29 September 2025.

Instrument validation

All analyses were performed in R 4.3.3, using the tidyverse packages [45] and the psych package [46].

Content & face validity. To determine content validity, we used the Content Validity Index (CVI) [41]. To calculate this index, participants' responses on the relevancy and representativeness of the items were dichotomised: not or somewhat relevant were coded as 0, whereas quite or highly relevant were coded as 1. The CVI was calculated for each item separately, and items had to reach a standard cutoff value of at least 0.80 to remain part of the scale [41]. Additionally, we calculated scale-wide average CVI (S-CVI/Ave), and the scale-wide CVI with universal agreement (S-CVI/UA). The S-CVI/Ave is a metric to determine the average CVI of the whole scale, whereas the S-CVI/UA indicates the percentage of items rated as relevant by all experts. Both indices were required to reach a standard minimum cutoff value of 0.80 [41]. If these cutoffs were not met, items with the lowest CVI values were iteratively removed until acceptable values were reached.

To determine face validity, we used the face validity index (FVI) [43]. This index is calculated analogously to the CVI based on end-users' ratings of item clarity and comprehensibility. We determined the FVI, the scale-wide average FVI (S-FVI/Ave), and the scale-wide FVI with universal agreement (S-FVI/UA). A cutoff of 0.80 or higher was considered acceptable for all validity metrics [43].

Construct validity

To determine construct validity, we measured the correlations of the TAIGHA scale with related questionnaires (to determine convergent validity), and with unrelated questionnaires (divergent validity). For convergent validity, we used the most widely used Trust in Automated Systems Survey [17], the Propensity to Trust in Technology Scale [47]. Because the AI may have given the same recommendation as participants' initial appraisal, we operationalized reliance as participants' shifts, i.e., included cases in which the AI's advice and participants' initial appraisal differed and determined whether or not they had changed their own appraisal in favor of the AI advice [48]. If participants' second appraisal aligned with the AI's advice, this was coded as reliance; if it did not, this was coded as no reliance. Cases in which participants' initial appraisal already aligned with the AI's recommendation were excluded from the analysis, because no shift toward the advice was possible.

Correlations above 0.30 were considered moderate and thus evidence of a related construct, whereas values above 0.90 were considered too high, as they would measure nearly the same construct and a new questionnaire would not be needed [49].

To determine divergent validity, we used the Reading Flow Short Scale [50], the General Self-Efficacy Short Scale [51], and the NASA Task Load Index (NASA-TLX) [52]. We specifically chose these questionnaires as they may be used to assess the use of DST, but there is no evidence that trust in AI-generated health advice should be correlated with reading flow, self-efficacy or task load. However, we excluded the self-rated performance and frustration subscales of the NASA-TLX, as these were expected to be at least somewhat related to trust and reliance on AI advice. Correlations below 0.30 were considered low and thus evidence of high divergent validity [49].

Criterion validity

To determine criterion validity, we measured the correlations of the TAIGHA scale with the outcome it aims to predict, that is, participants' reliance on the AI's health advice (also referred to as trust behaviour or advice-taking). We conceptualized reliance as the behavioural consequence of trust, where trust itself refers to the individual's subjective attitude toward the advice [24]. In the present study, reliance was operationalized as advice-taking, that is, whether participants changed their initial appraisal in favour of the AI-generated advice. Although reliance and advice-taking can be considered equivalent within this study, reliance as a behavioural outcome may be operationalized differently in other studies. Again, correlations above 0.30 were considered moderate and thus evidence of criterion validity [49].

Confirmatory factor analysis. We conducted a CFA using maximum likelihood estimation to assess whether the items show the same factorial structure as proposed by the underlying theory. Item-level missing data were handled using listwise deletion. The model fit was evaluated using commonly applied goodness-of-fit indices [34,53]: as incremental fit indices, we used the goodness of fit index (GFI), the comparative fit index (CFI), the Tucker-Lewis Index (TLI), and the Normed Fit Index (NFI). As absolute fit indices, we used the root mean square error of approximation (RMSEA), and the standardized root mean squared residual (SRMR). Additionally, we conducted an exploratory factor analysis using Pro max rotation. Because trust and distrust were conceptualized as related dimensions, we chose an oblique rotation to account for the expected correlation between these factors.

Reliability. Lastly, we determined the reliability of the TAIGHA scale using measures of internal consistency, i.e., the degree to which items within each subscale yield similar responses [54]. We calculated Cronbach's Alpha [55] and McDonald's Omega [56] to quantify internal consistency. Values above 0.75 were considered acceptable [57]. To assess item heterogeneity, we additionally calculated inter-item correlations.

Development of short scale

To develop the short version (TAIGHA-S), we followed the approach introduced in a previous article [58], which was originally based on the development of a short scale by Wessel et al. [59]. To make sure that the scale can be used without high time demands, we aimed to include two items per subscale [60]. Based on our underlying theory, we aimed to include one cognitive (dis-)trust and one affective (dis-)trust item each. To select items, we first used the item-total correlation of each item (measuring the correlation between each item and the total score when that item was excluded [61]) and selected items with the highest values. Second, when multiple items had identical item-total correlations, we used factor loadings from the CFA to select the candidate items that should be included. To validate the short scale, we correlated it with the full version and repeated the full validation process: assessing convergent and divergent validity, conducting a CFA and determining reliability.

Results

Sample characteristics

For the content validation, we surveyed 10 domain experts: eight (80%) were postdoctoral researchers, one (10%) an assistant professor, and one (10%) an associate or full professor. Three experts (30%) reported their primary field of expertise to be medicine and healthcare, three (30%) human-computer interaction, two (20%) psychology, and two (20%) AI or computer science. Four (40%) reported having conducted research on trust, six (60%) on human-computer interaction, seven (70%) on digital health, and seven (70%) on questionnaire development. For the face validation, we surveyed 30 participants who were exactly balanced in gender and on average 41 years old (SD = 14). For the psychometric evaluation, we surveyed 393 participants, of whom eight (2%) were excluded for answering attention check questions incorrectly. The characteristics of the final included sample are shown in Table 1.

Table 1. Characteristics of the sample included for the psychometric evaluation.

Characteristic	Total
Age (years), M(SD)	30 (16)
Gender, n (%)	
Male	184 (47.8%)
Female	199 (51.7%)
Other	1 (0.0%)
Prefer not to say	1 (0.0%)
Education, n (%)	
Finished high school with no qualifications	5 (1.3%)
Secondary school-leaving certificate	14 (3.6%)
High school diploma	50 (13.0%)
Completed apprenticeship	13 (3.4%)
Vocational secondary diploma	32 (8.3%)
A-Levels	38 (9.9%)
University degree	233 (60.5%)
Self-Efficacy, M(SD)[a]	4.0 (0.8)
Trust in Technology, M(SD)[a]	3.6 (0.5)

[a]On a 5-point Likert scale.

https://doi.org/10.1371/journal.pdig.0001488.t001

Item generation with AI-GENIE

Initially, 60 items were generated. The overall questionnaire had an NMI of 90.6% (60.0% for the trust subscale and 67.9% for the distrust subscale) and item stability values between 5% and 100%. This indicates poor psychometric properties. After the automated item selection, 28 items were left and the reduced questionnaire showed good psychometric properties with an NMI value of 100% (100% for both subscales) and item stability values between 81% and 100%. The automatically reduced questionnaire was then given to domain experts to further reduce the items to include only items that demonstrate high content validity.

Content validity

The CVI-I values of the reduced questionnaire ranged from 0.7 to 1.0, see Table 2. Because 8 items did not reach the cut-off of 0.8, those items were deleted. The S-CVI/Ave was 0.90 and therefore acceptable, but the S-CVI/UA was 0.45 and not acceptable. Following a conservative approach, items with a CVI-I of 0.8 were then deleted to improve content validity. This resulted in acceptable values of S-CVI/Ave = 0.99 and S-CVI/UA = 0.90.

Table 2. Content validity indices of the initial 28 items.

Item	Rated as not relevant (n)	Rated as relevant (n)	CVI-I
Initial Trust Item 1	3	7	0.70
Initial Trust Item 2	0	10	1.00
Initial Trust Item 3	0	10	1.00
Initial Trust Item 4	3	7	0.70
Initial Trust Item 5	2	8	0.80
Initial Trust Item 6	2	8	0.80
Initial Trust Item 7	0	10	1.00
Initial Trust Item 8	2	8	0.80
Initial Trust Item 9	3	7	0.70
Initial Trust Item 10	3	7	0.70
Initial Trust Item 11	2	8	0.80
Initial Trust Item 12	4	6	0.60
Initial Trust Item 13	0	10	1.00
Initial Trust Item 14	3	7	0.70
Initial Trust Item 15	1	9	0.90
Initial Trust Item 16	2	8	0.80
Initial Trust Item 17	2	8	0.80
Initial Distrust Item 1	2	8	0.80
Initial Distrust Item 2	0	10	1.00
Initial Distrust Item 3	2	8	0.80
Initial Distrust Item 4	0	10	1.00
Initial Distrust Item 5	2	8	0.80
Initial Distrust Item 6	0	10	1.00
Initial Distrust Item 7	2	8	0.80
Initial Distrust Item 8	0	10	1.00
Initial Distrust Item 9	3	7	0.70
Initial Distrust Item 10	3	7	0.70
Initial Distrust Item 11	0	10	0.10

https://doi.org/10.1371/journal.pdig.0001488.t002

Face validity

All FVI-I values of the included items were acceptable, ranging from 0.7 to 1.0, see Table 3. With S-CVI/Ave = 0.99 and S-CVI/AU = 0.80, scale-wide face validity was acceptable as well.

Construct validity. All items had few (or no) missing responses, low skewness and kurtosis; Table 4**.** The item difficulty was within acceptable values, and all items had high item-total correlation.

Convergent validity. The trust and distrust subscales showed a very high correlation (r = 0.67 [95% CI 0.61 to 0.72] and r = -0.66 [95% CI -0.60 to -0.71], respectively) with the Trust in Automated Systems Survey, as well as high correlations with the Propensity to Trust Scale (r = 0.53 [95% CI 0.46 to 0.60] and r = -0.46 [95% CI -0.38 to -0.53]), see Table 5.

Divergent validity. The trust and distrust subscales showed low to very low correlations with unrelated items. The highest correlation was between the NASA-TLX mental demand subscale and the distrust items (r = 0.25 [95% CI 0.15 to 0.34]), and the lowest correlation was between the trust subscale and the NASA-TLX effort subscale (r = 0.00 [95% CI -0.10 to 0.10]); see Table 6.

Table 3. Face validity indices of the items remaining after content validation.

Item	Rated as not clear and understandable (n)	Rated as clear and understandable (n)	FVI-I
Trust Items			
1: I feel comfortable about following the AI's health advice	0	30	1.00
2: The AI provides advice that should lead to better health.	1	29	0.97
3: I feel a sense of trust toward the AI's advice	0	30	1.00
4: I feel positive about relying on the AI's advice.	0	30	1.00
5: I believe the AI's advice will help me make good health decisions.	0	30	1.00
Distrust Items			
1: I feel uncomfortable with the idea of following the AI's advice.	0	30	1.00
2: I think following the AI's advice could pose health risks.	0	30	1.00
3: I feel distressed about relying on the AI's advice.	0	30	1.00
4: I feel tense about taking the AI's advice into account.	0	30	1.00
5: The idea of trusting this AI advice makes me uneasy.	2	28	0.93

https://doi.org/10.1371/journal.pdig.0001488.t003

Table 4. Descriptive statistics, item difficulty, and item-total correlation of the included items.

Item	Missing data (%)	Mean (SD)	Median (IQR)	Skewness	Kurtosis	Min - Max	Item Difficulty	Item-Total Correlation
Trust Item 1	0%	3.78 (0.90)	4 (3–4)	-0.9	0.9	1-5	0.76	0.84
Trust Item 2	0%	3.77 (0.76)	4 (3–4)	-0.6	0.6	1-5	0.75	0.78
Trust Item 3	0.3%	3.55 (0.89)	4 (3–4)	-0.9	0.6	1-5	0.71	0.85
Trust Item 4	0.3%	3.58 (0.92)	4 (3–4)	-0.7	0.2	1-5	0.72	0.85
Trust Item 5	0%	3.74 (0.81)	4 (3–4)	-0.9	1.3	1-5	0.75	0.85
Distrust Item 1	0%	2.32 (1.06)	2 (2–3)	0.8	-0.2	1-5	0.47	0.80
Distrust Item 2	0%	2.14 (0.95)	2 (2–3)	0.9	0.4	1-5	0.43	0.79
Distrust Item 3	0%	1.96 (0.94)	2 (1–2)	1.1	1.2	1-5	0.39	0.79
Distrust Item 4	0.3%	2.16 (1.05)	2 (12–3)	0.9	0.1	1-5	0.43	0.86
Distrust Item 5	0%	2.29 (1.07)	2 (2–3)	0.8	0.0	1-5	0.46	0.85

https://doi.org/10.1371/journal.pdig.0001488.t004

Table 5. Convergent validity. Correlations with other variables that measure related concepts.

	Trust Items	Distrust Items	Trust in Automated Systems Survey	Propensity to Trust Scale
Trust Items	1			
Distrust Items	-0.79	1		
Trust in Automated Systems Survey	0.67	-0.66	1	
Propensity to Trust Scale	0.53	-0.46	0.42	1

https://doi.org/10.1371/journal.pdig.0001488.t005

Table 6. Divergent validity. Correlations with other variables that measure unrelated concepts.

	Trust Items	Distrust Items	Reading Flow Short Scale	General Self-Efficacy Short Scale	(NASA-TLX)			
					Mental Demand	Physical Demand	Temporal Demand	Effort
Trust Items	1							
Distrust Items	-0.79	1						
Reading Flow Short Scale	0.19	-0.18	1					
General Self-Efficacy Short Scale	0.13	-0.11	0.23	1				
Mental Demand (NASA-TLX)	-0.14	0.25	-0.02	-0.10	1			
Physical Demand (NASA-TLX)	-0.12	0.20	-0.16	-0.15	0.48	1		
Temporal Demand (NASA-TLX)	-0.10	0.15	-0.22	-0.22	0.44	0.54	1	
Effort (NASA-TLX)	0.00	0.06	0.08	-0.02	0.43	0.21	0.19	1

https://doi.org/10.1371/journal.pdig.0001488.t006

Criterion validity. Criterion validity was assessed in the subsample of participants whose initial triage decision differed from the AI recommendation and who therefore had the opportunity to shift toward the AI advice ($n = 146$). In this subsample, the trust subscale showed a moderate correlation with reliance on AI advice ($r = 0.35$ [95% CI 0.20 to 0.49]), whereas the distrust subscale showed a low correlation with reliance on AI advice ($r = -0.19$ [95% CI -0.02 to -0.34]). Correlations of the trust subscale with reliance on AI advice were higher than the correlations of the Trust in Automated Systems Survey ($r = 0.24$ [95% CI 0.09 to 0.39]) and the Propensity to Trust Scale ($r = 0.07$ [95% CI -0.10 to 0.23]) with reliance on AI advice; see Table 7.

CFA

The CFA was conducted using complete cases with no missing responses ($n = 382$). The proposed two-factor model fit the data well, see Table 8 for goodness-of-fit metrics. All fit indices met the predefined cutoff values, including the RMSEA (0.07), which was below the threshold of 0.08. The remaining indices indicated strong fit (CFI = 0.98, TLI = 0.98, NFI = 0.97, SRMR = 0.03). The standardised factor loadings ranged from 0.80 to 0.90, and, as expected, both factors had a high correlation but were separate factors ($r = -0.84$), see Fig 2**.** Given the high correlation between these factors, we additionally conducted an exploratory factor analysis (which supported our two-factor structure) as well as a CFA with a one-factor model, which demonstrated worse fit than the two-factor structure ($\Delta\chi^2(1) = 125.27$, $p < .001$), as shown in Fig B in S3 Appendix, Table C in S3 Appendix and Table D in S3 Appendix.

Reliability

The internal consistency of the TAIGHA scale was excellent, with Cronbach's Alpha = 0.95 and McDonald's Omega = 0.96. The two subscales demonstrated high internal consistency as well ($\alpha = 0.94$ and $\omega = 0.94$ for the trust subscale, and $\alpha = 0.93$ and $\omega = 0.93$ for the distrust subscale). Inter-item correlations were high (mean $r = 0.75$, range: 0.70-0.79 for trust; mean $r = 0.73$, range = 0.64-0.81 for distrust), as shown in Table E in S3 Appendix and Table F in S3 Appendix.

Table 7. Criterion validity. Correlations with the outcome it aims to predict.

	Reliance on AI advice
Trust Items	0.35
Distrust Items	-0.19
Trust in Automated Systems Survey	0.24
Propensity to Trust Scale	0.07

https://doi.org/10.1371/journal.pdig.0001488.t007

Table 8. Fit Indices for the proposed two-factor model.

Fit index	Cutoff	Value
GFI	> 0.90	0.95
CFI	> 0.90	0.98
TLI	> 0.90	0.98
NFI	> 0.90	0.97
RMSEA	< 0.08	0.07
SRMR	< 0.08	0.03

https://doi.org/10.1371/journal.pdig.0001488.t008

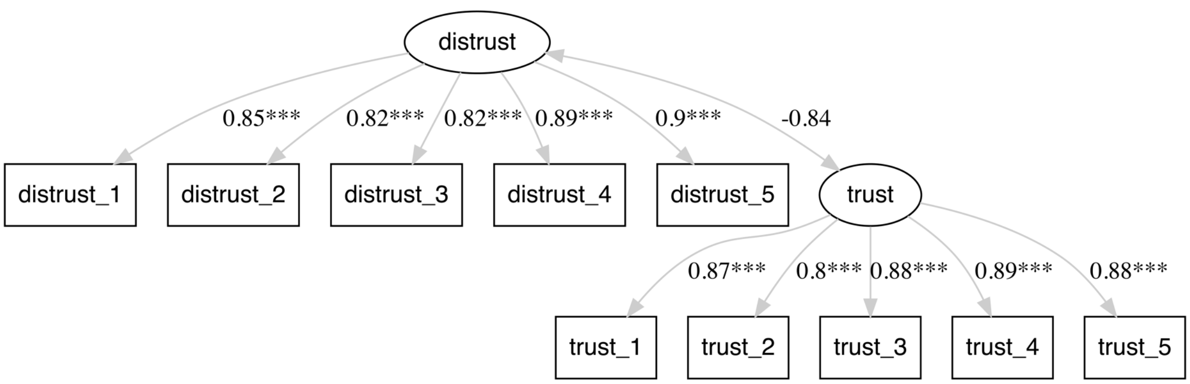

Fig 2. Factor structure and standardized factor loadings.

https://doi.org/10.1371/journal.pdig.0001488.g002

Short scale

For the trust subscale, item five had the highest ITC among cognitive trust items, and items three and four had the highest ITC among affective trust items (ITC = 0.85 for all three items). Item four had a slightly higher factor loading (0.89) than item three (0.88) and was therefore chosen for the short scale. For the distrust subscale, item two had the highest ITC among cognitive distrust items (ITC = 0.79), and item four had the highest ITC among affective distrust items (ITC = 0.86). The TAIGHA-S short scale thus consists of items four and five for the trust subscale and items two and four for the distrust subscale.

The short scale and its subscales showed a high correlation with the full scale and subscales (Table 9). It also demonstrated high convergent validity through high correlations with related concepts, high criterion validity through a moderate correlation with reliance on AI advice (Table 9) and high divergent validity through low correlations with unrelated concepts (Table 10).

Similar to the full scale, the CFA demonstrated a good fit (using n = 383 complete cases after listwise deletion), with factor loadings ranging between 0.84 and 0.90 (Fig 3), and all goodness-of-fit metrics reaching cutoff values (Table 11).

Table 9. Convergent and criterion validity. Correlations with other variables that measure related concepts and with the outcome it aims to predict.

Unrelated Construct	Trust Short Subscale	Distrust Short Subscale
Full Trust Items	0.96	-0.73
Full Distrust Items	-0.76	0.96
Trust in Automated Systems Survey	0.63	-0.61
Propensity to Trust Scale	0.51	-0.40
Reliance on AI Advice	0.35	-0.14

https://doi.org/10.1371/journal.pdig.0001488.t009

Table 10. Divergent validity. Correlations with other variables that measure unrelated concepts.

Unrelated Construct	Trust Short Subscale	Distrust Short Subscale
Reading Flow Short Scale	0.17	-0.17
General Self-Efficacy Short Scale	0.11	-0.10
Mental Demand (NASA-TLX)	-0.15	0.25
Physical Demand (NASA-TLX)	-0.12	0.22
Temporal Demand (NASA-TLX)	-0.06	0.16
Effort (NASA-TLX)	-0.02	0.06

https://doi.org/10.1371/journal.pdig.0001488.t010

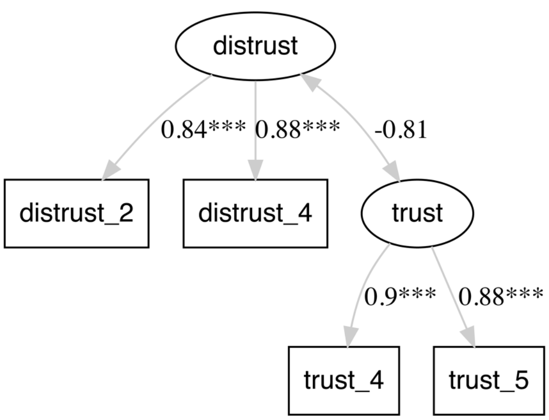

Fig 3. Factor structure and standardized factor loadings of the short scale.

https://doi.org/10.1371/journal.pdig.0001488.g003

The reliability of the short scale was acceptable (α = 0.88 and ω = 0.93 for the full scale, α = 0.88 and ω = 0.89 for the trust subscale, and α = 0.84 and ω = 0.85 for the distrust subscale).

Discussion

Principal findings

This study aimed to develop and validate a measurement instrument that specifically assesses trust in AI-generated health advice. Given the limitations of existing questionnaires that measure only general trust in technology, lack contextual sensitivity and assess perceived trustworthiness of systems rather than the user's willingness to follow a specific piece of health advice generated by an AI tool [17,24], we developed and validated the TAIGHA scale and its short form, TAIGHA-S, which can be found in S1 and S2 Appendices.

Table 11. Fit Indices for the short scale.

Fit index	Cutoff	Value
GFI	> 0.90	0.99
CFI	> 0.90	1.00
TLI	> 0.90	1.00
NFI	> 0.90	0.99
RMSEA	< 0.08	0.00
SRMR	< 0.08	0.00

https://doi.org/10.1371/journal.pdig.0001488.t011

Both scales were explicitly designed for AI-generated health advice and are based on empirical findings from previous trust scales and McAllister's trust theory, that is, they distinguish between trust and distrust as highly related but separate constructs [20,25], considering both cognitive and affective items [16]. The TAIGHA scale demonstrated high content validity as rated by experts from multiple relevant domains and was found to be face valid, that is, understandable to a lay audience. It also showed excellent reliability, very high goodness-of-fit indices for a two-factor model, and high convergent and divergent validity. Thus, the TAIGHA scale appears to measure users' trust in AI-generated health advice reliably without capturing unrelated constructs. The high reliability should be interpreted with respect to the scale's intended scope: it was designed to measure cognitive and affective components of trust and distrust in a brief format relevant to reliance on AI-generated health advice. The high inter-item correlations are consistent with that focus, as they suggest a narrow measurement (albeit with limited construct breadth) rather than a measurement of all possible trust facets [62]. Notably, in our criterion validation, the TAIGHA scale correlated more strongly with whether participants followed the AI's advice than the existing general-purpose Trust in Automated Systems Survey [17]. More broadly, this finding suggests that a scale developed for a specific health advice-taking context can explain behaviour in that context more accurately than a general technology trust scale that lacks items adapted to the use case.

Our findings extend and combine several lines of research. First, the TAIGHA scale provides an example of how to operationalize state trust rather than trustworthiness - two concepts that many prior studies have conflated [24]. Much of the literature on trust in technology measures perceived trustworthiness attributes of the system (such as reliability or competence of a technology) instead of the user's state of trusting and acting on the advice it provides [14,24]. To address this, we applied McAllister's theoretical trust model [16], which focuses on cognitive and affective components within the trustor (lay users, in this case), rather than Mayer et al.'s model, which emphasises trustworthiness and characteristics of the trustee such as competence, integrity, and benevolence [15,63]. This conceptual difference likely explains why the TAIGHA scale correlated more strongly with reliance than the more general Trust in Automated Systems Survey, which is based on Mayer et al.'s model. Such a focus is especially relevant in health contexts, where AI may sometimes not be perceived as trustworthy, yet users still follow its advice [8].

Second, the TAIGHA scale treats distrust as a separate but related complementary dimension rather than the mere opposite of trust. Re-analyses of previous trust scales have repeatedly shown that complementary but separate trust and distrust factors better explain the factor structure of trust questionnaires than a single unidimensional construct [20,25]. Our results are consistent with this view and may suggest that interventions aimed at reducing distrust (e.g., by citing sources or providing explanations) may not necessarily increase trust. Conversely, interventions designed to increase trust may not automatically reduce users' scepticism or caution toward AI tools. Evaluating both subscales therefore enables researchers and developers to determine which component an intervention primarily affects or should be addressed.

Third, we employed a relatively new methodological approach that integrates generative AI into psychometrics [28,64–66]. Using the AI-GENIE framework, we first generated an item pool with GPT-4.1 and then selected items that demonstrated strong psychometric properties [28]. Prior research indicates that items generated by LLMs are often of higher quality than those written by humans [33], and that the automated validation procedure can produce items that also perform well in traditional empirical evaluations. Our results provide further evidence for the latter, as the automated validation process was technically successful. However, subsequent content validation by domain experts seems to remain important, as several psychometrically sound items in this study were removed for being irrelevant or not fitting to theory. This may be the result of LLMs' high output variability and their limited adherence to instructions [67–69]. Future research using the AI-GENIE approach should therefore iteratively refine prompts to ensure that the generated items fit well with the underlying theory and include a manual item selection phase to remove irrelevant items. Although such an automated approach can substantially accelerate item generation and validation, our results underscore that a manual construct and content validation remain important to guarantee the quality and meaningfulness of the scale items.

From a theoretical perspective, our findings extend the two-factor perspective (trust and distrust) from general trust theory [16] and general trust in technology research [17,20,25] to a domain-specific state trust measurement. Second, they reinforce the argument that trust in AI advice among lay users should be behaviour-proximal [24], that is, scales should ask users about their cognition and affect in the actual decision context rather than about the AI characteristics and their perceived trustworthiness, which was shown to improve correlation with reliance. Third, the affective components of trust, which have often been excluded from traditional technology trust scales [17,23], have been shown to be important predictors of trusting behaviour, which may provide new opportunities for trust interventions and AI design.

From a practical standpoint, the TAIGHA scale offers a highly relevant outcome measure for evaluating public-facing AI tools. Although the instrument does not measure trust in an AI tool as a whole (and does not include aspects of interface design, institutional framing, regulatory cues, or perceived clinical authority), it was designed to measure users' state trust and distrust toward a specific advice instance presented in a controlled context. This makes the scale particularly suitable for studies that compare how advice is presented (e.g., explanation formats, uncertainty communication, or interface designs). When trust is included only as a secondary outcome, the short form, TAIGHA-S, can be used as an alternative to self-constructed single-item measures, as it has low resource demands but good psychometric properties [60]. Reporting trust and distrust separately also supports safety evaluations: in situations in which users rely excessively on low-quality or high-risk advice or rely insufficiently on high-quality advice, the scale may help identify whether this behaviour is associated with trust, distrust, or both. Treating trust and distrust as separate constructs is also practically relevant for the design and evaluation of AI-supported health communication, because interventions may affect them differently. For example, providing uncertainty information or references for claims may reduce distrust without necessarily increasing trust, whereas changes that make the advice appear more reassuring or personalized may increase trust without reducing safety concerns [9,70]. Measuring both dimensions separately in a two-factor model can therefore help researchers and developers both to design interventions and test whether they primarily increase trust, reduce distrust, or affect both dimensions in ways that may be relevant to safer and more appropriate reliance on AI-generated health advice.

By capturing trust and distrust as complementary constructs, the TAIGHA scales may help researchers test and identify factors that influence users' behaviour. Accordingly, researchers can then design interventions that strengthen state trust in reliable AI outputs and/or mitigate state distrust that limits appropriate use. Furthermore, validated tools such as TAIGHA-S provide a low-resource yet psychometrically robust method for evaluating these perceptions, enabling efficient assessment of AI tools in healthcare and supporting their safe and effective adoption. This is particularly important in contexts where the growing use of large language model (LLM) technologies serves as a source of healthcare guidance for lay users, particularly in light of the persistent resource constraints faced by healthcare systems. At an organizational level, the TAIGHA scales may also support the responsible implementation of AI in health care. By distinguishing between trust and distrust, they can help healthcare organizations assess whether AI-supported communication promotes appropriate trust rather than either uncritical acceptance or inappropriate rejection of advice. This is also relevant for the application of global health policies and ethical guidelines for AI implementation, which often focus on transparency, explainability, and potential deployment risks [71]. When used with behavioural and clinical outcomes, the TAIGHA scales may therefore support the responsible evaluation of implementation strategies for AI-generated health advice.

Limitations

The development and validation of the TAIGHA scale have several limitations. Although we used a relatively high sample size of 385 participants [44,72], the principal limitation of this study is that the online-survey format with UK participants may have resulted in a sample with greater technological affinity, digital literacy, and familiarity with online health information than the general UK population. Accordingly, the present findings should be interpreted as an initial validation in an English-speaking sample of people living in the UK. In addition, the trust items showed a ceiling-adjacent distribution, with means between 3.55 and 3.78 on a 5-point Likert scale and negative skewness. This may reflect high trust in AI-generated health advice in the Prolific online sample, but it may also limit the accurate measurement of higher trust levels. Future studies should examine the scale in different populations, including groups with very positive and very negative attitudes toward AI, lower digital health literacy, lower socioeconomic status, and less familiarity with AI DSTs, to determine whether the scale performs similarly across these groups and supports equitable clinical applicability [73]. In addition, the present study does not determine whether the TAIGHA scale can be used across different cultural contexts and languages. Before international use, the scale should undergo cross-cultural validation in other countries, including testing measurement invariance to determine whether the underlying factor structure remains comparable across populations. Future studies should also translate the TAIGHA scales into other languages, apply cultural adaptation where needed, and re-validate these versions. We further did not conduct differential item functioning analyses by demographic variables. Although the overall sample size was appropriate for the psychometric validation, the study was not powered for stable subgroup-specific DIF estimation (particularly for smaller education categories). This limits conclusions about whether the items function equivalently across demographic groups. Future validation studies should therefore examine DIF and measurement invariance by gender, age, education level, and other equity-relevant variables in larger and more diverse samples.

Another limitation is the mode of interaction. Because participants completed the TAIGHA scale after a single exposure to one instance of AI-generated health advice, the present validation only measures state trust and distrust, but not longitudinal trust calibration. In real-world use, trust in AI tools may change over repeated interactions as users gain experience with the AI DST [14]. Although this may constrain generalisability to long-term interactions, in many scenarios such as assessing acute symptoms, users typically interact only once with an AI DST rather than repeatedly [74]. Future studies should examine the applicability and validity of the TAIGHA scale, particularly its convergent validity and its association with reliance in longitudinal or repeated-use settings. For example, participants could complete the scale after multiple advice instances to examine whether TAIGHA scores change with repeated use, and whether these changes correspond to the AI DST's accuracy. Such studies would help determine whether TAIGHA is applicable for long-term use or whether additional items or a different underlying theory may be needed.

A further limitation is that the present validation did not include any sociotechnical influences that impact trust in real-world deployment. The TAIGHA scale was specifically developed for measuring state trust and validated in a controlled scenario in which such influences were minimized. Future studies should develop measurements of trait trust to allow measuring an individual's disposition to trust AI-generated health advice.

Similarly, another limitation relates to our criterion validation. We operationalized reliance as a shift in self-triage level toward the AI recommendation. Although this provides an easy-to-interpret behavioural proxy, it does not measure more implicit forms of reliance, such as seeking verification from other sources, delaying action, or selectively adopting only parts of the advice. The modest association between the TAIGHA scores and reliance may therefore – at least partly – reflect this constrained proxy for behaviour. Future studies should validate the scale against additional endpoints, such as changes in decisional certainty, information verification behaviour, or selective adherence to different aspects of the given advice.

We acknowledge that using scenarios made the empirical validation somewhat artificial, as these are not real-world interactions. However, this approach represents an appropriate trade-off between data quality and ethical feasibility, as asking participants experiencing real acute symptoms to use an AI tool and answer multiple questions would raise ethical concerns. To approximate real-world behaviour, the use case scenarios and experimental setup in this study were validated in previous research, included real patient cases, and were shown to be representative with respect to the use case [39,40]. Nevertheless, it would be valuable to further validate the TAIGHA scale in future studies using different health advice scenarios, such as those involving self-care or health promotion. Beyond scenario-based validation, future studies should evaluate the TAIGHA scales in prospective real-world studies and clinical trials in which patients receive and act on AI-generated health advice in actual care contexts. This would allow testing whether the scales also have predictive validity regarding actual patient adherence, decision-making and clinically relevant health outcomes. Also, the scale focused on assessing trust in AI by lay users. Future studies should explore to what extent this scale may also be extended to assess trust in expert-facing AI tools. The high internal consistency and inter-item correlations suggest high reliability but limited construct breadth. The TAIGHA scale thus appears well suited for brief assessment of trust and distrust according to our underlying theory, but it may be less sensitive to broader subfacets of these constructs. Although a broad measurement was not the aim of the present instrument, future scales could include supplemental items to measure additional aspects of trust and distrust.

Although the TAIGHA scale showed a moderate correlation with reliance and a stronger association with reliance than the Trust in Automated Systems Survey, its correlation with and thus predictive power for actual behaviour remains limited. This outcome is expected, as decisions are also influenced by additional factors, such as individuals' heuristics when taking advice into account, their integration of the advice with pre-existing knowledge and experience, and situational circumstances [9,14,75,76]. Therefore, researchers should, whenever possible, include measures of actual behaviour alongside trust assessments.

Lastly, the study was not preregistered. Although preregistration is less common in instrument development, it would have increased transparency regarding which analytic decisions were planned a priori.

Conclusions

In this article, we introduced and validated the TAIGHA scale and its four-item short form (TAIGHA-S) to assess users' trust and distrust in AI-generated health advice. Both instruments demonstrated excellent psychometric properties and correlated more strongly with reliance than a generic technology trust scale which indicates added value for health-advice contexts. The TAIGHA and TAIGHA-S can serve as outcome measures for evaluating public-facing, and potentially clinician-facing AI systems. Reporting trust and distrust as separate constructs allows researchers to identify whether an intervention increases acceptance of AI advice, reduces aversion to it, or both. Overall, the TAIGHA scales provide a theory-grounded, use-case-specific measure of state trust that can help identify when people accept or reject AI-generated health advice. Future studies should translate the TAIGHA scales into different languages and validate the scales in prospective real-world and clinical studies to examine whether they predict actual patient behaviour.

Supporting information

S1 Appendix. Trust in AI generated health advice (TAIGHA) scale.
(PDF)

S2 Appendix. Trust in AI generated health advice short (TAIGHA-S) scale.
(PDF)

S3 Appendix. Additional information and analyses.
(DOCX)

Author contributions

Conceptualization: Marvin Kopka, Markus Feufel.

Data curation: Marvin Kopka.

Formal analysis: Marvin Kopka.

Investigation: Marvin Kopka.

Methodology: Marvin Kopka.

Project administration: Marvin Kopka, Austen El-Osta.

Resources: Marvin Kopka, Azeem Majeed, Austen El-Osta, Markus Feufel.

Software: Marvin Kopka.

Supervision: Austen El-Osta, Markus Feufel.

Validation: Gabriella Spinelli, Austen El-Osta, Markus Feufel.

Visualization: Marvin Kopka.

Writing – original draft: Marvin Kopka.

Writing – review & editing: Marvin Kopka, Azeem Majeed, Gabriella Spinelli, Austen El-Osta, Markus Feufel.

References

1. Chatterji A, Cunningham T, Deming D, Hitzig Z, Ong C, Shan CY. How People Use ChatGPT. Cambridge, MA: National Bureau of Economic Research. 2025. https://doi.org/10.3386/w34255
2. Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. Bull World Health Organ. 2020;98(4):251–6. https://doi.org/10.2471/BLT.19.237487 PMID: 32284648
3. Chustecki M. Benefits and Risks of AI in Health Care: Narrative Review. Interact J Med Res. 2024;13:e53616. https://doi.org/10.2196/53616 PMID: 39556817
4. Sorich MJ, Menz BD, Hopkins AM. Quality and safety of artificial intelligence generated health information. BMJ. 2024;:q596. https://doi.org/10.1136/bmj.q596
5. Eichenberger A, Thielke S, Van Buskirk A. A Case of Bromism Influenced by Use of Artificial Intelligence. AIM Clinical Cases. 2025;4(8). https://doi.org/10.7326/aimcc.2024.1260
6. Biswas SS. Role of Chat GPT in Public Health. Ann Biomed Eng. 2023;51(5):868–9. https://doi.org/10.1007/s10439-023-03172-7
7. Joseph J, Jose B, Jose J. The generative illusion: how ChatGPT-like AI tools could reinforce misinformation and mistrust in public health communication. Front Public Health. 2025;13:1683498. https://doi.org/10.3389/fpubh.2025.1683498 PMID: 41080885
8. Shekar S, Pataranutaporn P, Sarabu C, Cecchi GA, Maes P. People overtrust AI-generated medical advice despite low accuracy. NEJM AI. 2025;2(6):AIoa2300015. https://doi.org/10.1056/AIoa2300015
9. Kopka M, Wang SM, Kunz S, Schmid C, Feufel MA. Technology-supported self-triage decision making. npj Health Syst. 2025;2(1). https://doi.org/10.1038/s44401-024-00008-x
10. Kopka M, Schmieding ML, Rieger T, Roesler E, Balzer F, Feufel MA. Determinants of Laypersons' Trust in Medical Decision Aids: Randomized Controlled Trial. JMIR Hum Factors. 2022;9(2):e35219. https://doi.org/10.2196/35219 PMID: 35503248
11. Eghrari AO. Artificial Intelligence, Medical Knowledge, and Empowering Patients. Mayo Clinic Proceedings: Digital Health. 2024;2(1):160–2. https://doi.org/10.1016/j.mcpdig.2024.01.008
12. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ. 2024;58(11):1276–85. https://doi.org/10.1111/medu.15402 PMID: 38639098
13. Kahr PK, Rooks G, Willemsen MC, Snijders CCP. Understanding Trust and Reliance Development in AI Advice: Assessing Model Accuracy, Model Explanations, and Experiences from Previous Interactions. ACM Trans Interact Intell Syst. 2024;14(4):1–30. https://doi.org/10.1145/3686164
14. Hoff KA, Bashir M. Trust in automation: integrating empirical evidence on factors that influence trust. Hum Factors. 2015;57(3):407–34. https://doi.org/10.1177/0018720814547570 PMID: 25875432
15. Mayer RC, Davis JH, Schoorman FD. An integrative model of organizational trust. AMR. 1995;20(3):709–34. https://doi.org/10.5465/amr.1995.9508080335

16. McAllister DJ. Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. Academy of Management Journal. 1995;38(1):24–59. https://doi.org/10.2307/256727

17. Jian J-Y, Bisantz AM, Drury CG. Foundations for an Empirically Determined Scale of Trust in Automated Systems. International Journal of Cognitive Ergonomics. 2000;4(1):53–71. https://doi.org/10.1207/s15327566ijce0401_04

18. Gutzwiller RS, Yousefi R, Larson-Calcano T, Lee JR, Verma A, Tenhundfeld NL, et al. Systematic Review of the Use and Modification of the "Trust in Automated Systems Scale". Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2025;69(1):2176–81. https://doi.org/10.1177/10711813251357911

19. Gutzwiller RS, Chiou EK, Craig SD, Lewis CM, Lematta GJ, Hsiung CP. Positive Bias in the 'Trust in Automated Systems Survey'? An Examination of the Jian *et al*. (2000) Scale. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 2019. 217–21. https://doi.org/10.1177/1071181319631201

20. Spain RD, Bustamante EA, Bliss JP. Towards an Empirically Developed Scale for System Trust: Take Two. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2008;52(19):1335–9. https://doi.org/10.1177/154193120805201907

21. Gulati S, Sousa S, Lamas D. Design, development and evaluation of a human-computer trust scale. Behaviour & Information Technology. 2019;38(10):1004–15. https://doi.org/10.1080/0144929x.2019.1656779

22. Shang R, Hsieh G, Shah C. Trusting your AI agent emotionally and cognitively: Development and validation of a semantic differential scale for AI trust. AIES. 2024;7:1343–56. https://doi.org/10.1609/aies.v7i1.31728

23. Hoffman RR, Mueller ST, Klein G, Litman J. Measures for explainable AI: explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. Frontiers in Computer Science. 2023;5. https://doi.org/10.3389/fcomp.2023.1096257

24. Schlicker N, Baum K, Uhde A, Sterz S, Hirsch MC, Langer M. How do we assess the trustworthiness of AI? Introducing the trustworthiness assessment model (TrAM). Computers in Human Behavior. 2025;170:108671. https://doi.org/10.1016/j.chb.2025.108671

25. Scharowski N, Perrig SAC, von Felten N, Aeschbach LF, Opwis K, Wintersberger P, et al. To Trust or Distrust AI: A Questionnaire Validation Study. In: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, 2025. 361–74. https://doi.org/10.1145/3715275.3732025

26. Allen MS, Iliescu D, Greiff S. Single Item Measures in Psychological Science. European Journal of Psychological Assessment. 2022;38(1):1–5. https://doi.org/10.1027/1015-5759/a000699

27. Kohn SC, Kluck M, Shaw T. A Brief Review of Frequently Used Self-Report Measures of Trust in Automation. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2020;64(1):1436–40. https://doi.org/10.1177/1071181320641342

28. Russell-Lasalandra LL, Christensen AP, Golino H. Generative Psychometrics via AI-GENIE: Automatic Item Generation with Network-Integrated Evaluation. PsyArXiv. 2025. https://osf.io/fgbj4_v2

29. Golino HF, Epskamp S. Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. PLoS One. 2017;12(6):e0174035. https://doi.org/10.1371/journal.pone.0174035 PMID: 28594839

30. Danon L, Díaz-Guilera A, Duch J, Arenas A. Comparing community structure identification. J Stat Mech. 2005;2005(09):P09008. https://doi.org/10.1088/1742-5468/2005/09/P09008

31. Christensen AP, Garrido LE, Golino H. Unique Variable Analysis: A Network Psychometrics Method to Detect Local Dependence. Multivariate Behav Res. 2023;58(6):1165–82. https://doi.org/10.1080/00273171.2023.2194606 PMID: 37139938

32. OpenAI. GPT-4.1 Smartest non-reasoning model. https://platform.openai.com/docs/models/gpt-4.1

33. Martin Kowal J, Hurley Bryant K, Segall D, Kantrowitz T. Harnessing Generative AI for Assessment Item Development: Comparing AI-Generated and Human-Authored Items. Int J Selection Assessment. 2025;33(3). https://doi.org/10.1111/ijsa.70021

34. Yusoff MSB, Arifin WN, Hadie SNH. ABC of Questionnaire Development and Validation for Survey Research. EIMJ. 2021;13(1):97–108. https://doi.org/10.21315/eimj2021.13.1.10

35. Kopka M, Scatturin L, Napierala H, Fürstenau D, Feufel MA, Balzer F, et al. Characteristics of Users and Nonusers of Symptom Checkers in Germany: Cross-Sectional Survey Study. J Med Internet Res. 2023;25:e46231. https://doi.org/10.2196/46231 PMID: 37338970

36. Simpson RM, Jacques RM, Nicholl J, Stone T, Turner J. Measuring the impact introducing NHS 111 online had on the NHS 111 telephone service and the wider NHS urgent care system: an observational study. BMJ Open. 2022;12(7):e058964. https://doi.org/10.1136/bmjopen-2021-058964 PMID: 35820752

37. Wetzel A-J, Koch R, Koch N, Klemmt M, Müller R, Preiser C, et al. "Better see a doctor?" Status quo of symptom checker apps in Germany: A cross-sectional survey with a mixed-methods design (CHECK.APP). Digit Health. 2024;10:20552076241231555. https://doi.org/10.1177/20552076241231555 PMID: 38434790

38. El-Osta A, Riboli-Sasco E, Al Ammouri M, Altalib S, Luisa Neves A, Majeed A, et al. Factors Influencing the Use of Online Symptom Checkers in the United Kingdom: Cross-Sectional Study. JMIR Form Res. 2025;9:e65314. https://doi.org/10.2196/65314 PMID: 40953283

39. Kopka M, Napierala H, Privoznik M, Sapunova D, Zhang S, Feufel MA. The RepVig framework for designing use-case specific representative vignettes and evaluating triage accuracy of laypeople and symptom assessment applications. Sci Rep. 2024;14(1):30614. https://doi.org/10.1038/s41598-024-83844-z

40. Kopka M, Feufel MA. Statistical refinement of patient-centered case vignettes for digital health research. Front Digit Health. 2024;6:1411924. https://doi.org/10.3389/fdgth.2024.1411924 PMID: 39498100
41. Yusoff MSB. ABC of content validation and content validity index calculation. EIMJ. 2019;11(2):49–54. https://doi.org/10.21315/eimj2019.11.2.6
42. Palan S, Schitter C. Prolific.ac—A subject pool for online experiments. Journal of Behavioral and Experimental Finance. 2018;17:22–7. https://doi.org/10.1016/j.jbef.2017.12.004
43. Yusoff MSB. ABC of response process validation and face validity index calculation. EIMJ. 2019;11(3):55–61. https://doi.org/10.21315/eimj2019.11.3.6
44. Comrey AL, Lee HB. A first course in factor analysis. 2nd ed. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc. 1992.
45. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. JOSS. 2019;4(43):1686. https://doi.org/10.21105/joss.01686
46. Revelle W. Psych: Procedures for Psychological, Psychometric, and Personality Research. 2022. https://CRAN.R-project.org/package=psych
47. Jessup SA, Schneider TR, Alarcon GM, Ryan TJ, Capiola A. The Measurement of the Propensity to Trust Automation. Lecture Notes in Computer Science. Springer International Publishing. 2019. p. 476–89. https://doi.org/10.1007/978-3-030-21565-1_32
48. Bonaccio S, Dalal RS. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. Organizational Behavior and Human Decision Processes. 2006;101(2):127–51. https://doi.org/10.1016/j.obhdp.2006.07.001
49. Overholser BR, Sowinski KM. Biostatistics primer: part 2. Nutr Clin Pract. 2008;23(1):76–84. https://doi.org/10.1177/011542650802300176
50. Thissen BAK, Menninghaus W, Schlotz W. Measuring optimal reading experiences: The reading flow short scale. Front Psychol. 2018;9. https://doi.org/10.3389/fpsyg.2018.02542
51. Romppel M, Herrmann-Lingen C, Wachter R, Edelmann F, Düngen H-D, Pieske B, et al. A short form of the General Self-Efficacy Scale (GSE-6): Development, psychometric properties and validity in an intercultural non-clinical sample and a sample of patients at risk for heart failure. Psychosoc Med. 2013;10:Doc01. https://doi.org/10.3205/psm000091 PMID: 23429426
52. Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Hancock PA, Meshkati N. Advances in Psychology. North-Holland. 1988. 139–83. https://doi.org/10.1016/S0166-4115(08)62386-9
53. Goretzko D, Siemund K, Sterner P. Evaluating Model Fit of Measurement Models in Confirmatory Factor Analysis. Educ Psychol Meas. 2024;84(1):123–44. https://doi.org/10.1177/00131644231163813 PMID: 38250508
54. Streiner DL. Starting at the beginning: an introduction to coefficient alpha and internal consistency. J Pers Assess. 2003;80(1):99–103. https://doi.org/10.1207/S15327752JPA8001_18 PMID: 12584072
55. Cronbach LJ. Coefficient Alpha and the Internal Structure of Tests. Psychometrika. 1951;16(3):297–334. https://doi.org/10.1007/bf02310555
56. Hayes AF, Coutts JJ. Use Omega Rather than Cronbach's Alpha for Estimating Reliability. But…. Communication Methods and Measures. 2020;14(1):1–24. https://doi.org/10.1080/19312458.2020.1718629
57. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ. 2011;2:53–5. https://doi.org/10.5116/ijme.4dfb.8dfd PMID: 28029643
58. Kopka M, Slagman A, Schorr C, Krampe H, Altendorf M, Balzer F, et al. German mHealth App Usability Questionnaire (G-MAUQ) and short version (G-MAUQ-S): Translation and validation study. Smart Health. 2024;34:100517. https://doi.org/10.1016/j.smhl.2024.100517
59. Wessel D, Attig C, Franke T. ATI-S - An Ultra-Short Scale for Assessing Affinity for Technology Interaction in User Studies. In: Proceedings of Mensch und Computer 2019, 2019. 147–54. https://doi.org/10.1145/3340764.3340766
60. Eisinga R, Grotenhuis M te, Pelzer B. The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown?. Int J Public Health. 2013;58(4):637–42. https://doi.org/10.1007/s00038-012-0416-3 PMID: 23089674
61. Moosbrugger H, Kelava A. Deskriptivstatistische Evaluation von Items (Itemanalyse) und Testwertverteilungen. Testtheorie und Fragebogenkonstruktion. Berlin, Heidelberg: Springer Berlin Heidelberg. 2012. 75–102. https://doi.org/10.1007/978-3-642-20072-4_4
62. Clark LA, Watson D. Constructing validity: New developments in creating objective measuring instruments. Psychol Assess. 2019;31(12):1412–27. https://doi.org/10.1037/pas0000626 PMID: 30896212
63. Lee JD, See KA. Trust in automation: designing for appropriate reliance. Hum Factors. 2004;46(1):50–80. https://doi.org/10.1518/hfes.46.1.50_30392 PMID: 15151155
64. Barends AJ, de Vries RE. Developing and Improving Personality Inventories Using Generative Artificial Intelligence: The Psychometric Properties of a Short HEXACO Scale Developed Using ChatGPT 4.0. J Pers Assess. 2025;107(4):419–25. https://doi.org/10.1080/00223891.2024.2444454 PMID: 39727339
65. Ye H, Xie Y, Ren Y, Fang H, Zhang X, Song G. Measuring Human and AI Values Based on Generative Psychometrics with Large Language Models. AAAI. 2025;39(25):26400–8. https://doi.org/10.1609/aaai.v39i25.34839
66. Terry J, Strait G, Alsarraf S, Weinmann E, Waychoff A. Artificial intelligence in scale development: evaluating AI-generated survey items against gold standard measures. Current Psychology. 2025. https://doi.org/10.1007/s12144-025-08240-w
67. Atil B, Aykent S, Chittams A, Fu L, Passonneau RJ, Radcliffe E. Non-determinism of 'deterministic' LLM settings. arXiv. 2025. https://doi.org/10.48550/arXiv.2408.04667

68. Li Z, Peng B, He P, Yan X. Evaluating the Instruction-Following Robustness of Large Language Models to Prompt Injection. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024. 557–68. https://doi.org/10.18653/v1/2024.emnlp-main.33
69. Qin Y, Song K, Hu Y, Yao W, Cho S, Wang X, et al. InFoBench: Evaluating Instruction Following Ability in Large Language Models. In: Findings of the Association for Computational Linguistics ACL 2024, 2024. 13025–48. https://doi.org/10.18653/v1/2024.findings-acl.772
70. Meincke L, Nave G, Terwiesch C. Advice quality and source disclosure shape trust in AI-generated ethical advice. Sci Rep. 2026. https://doi.org/10.1038/s41598-026-44258-1
71. Ethics and governance of artificial intelligence for health: WHO guidance. 1st ed. Geneva: World Health Organization. 2021.
72. Anthoine E, Moret L, Regnault A, Sébille V, Hardouin J-B. Sample size used to validate a scale: a review of publications on newly-developed patient reported outcomes measures. Health Qual Life Outcomes. 2014;12:176. https://doi.org/10.1186/s12955-014-0176-2 PMID: 25492701
73. Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care — Addressing Ethical Challenges. N Engl J Med. 2018;378(11):981–3. https://doi.org/10.1056/NEJMp1714229
74. Meyer AND, Giardina TD, Spitzmueller C, Shahid U, Scott T, Singh H. Patient perspectives on the usefulness of an artificial intelligence-assisted symptom checker: cross-sectional survey study. J Med Internet Res. 2020;22(1):e14679. https://doi.org/10.2196/14679
75. Gigerenzer G, Todd PM. Fast and frugal heuristics: The adaptive toolbox. In: Simple heuristics that make us smart. New York, NY, US: Oxford University Press. 1999. 3–34
76. Klein G. Naturalistic decision making. Hum Factors. 2008;50(3):456–60. https://doi.org/10.1518/001872008X288385 PMID: 18689053