Journal of Clinical Epidemiology 187 (2025) 111962

OTHER GRADE PAPER

Trustworthiness of treatment clinical practice guidelines has modestly improved since the introduction of Institute of Medicine standards: a systematic survey

Maryam Ghadimi[a,*], Gordon Guyatt[a,b,c], Carlos Zaror[a,d], João Pedro Lima[a,e], Sana Gupta[f], Hamed Movahed[a], Wimonchat Tangamornsuksan[a], Gonzalo Bravo-Soto[a,g], Sahrish Masood[b], Huda Gomaa[h], Yanjiao Shen[a,i], Nima Behravan[b], Rachel Couban[e], Romina Brignardello-Petersen[a]

[a]*Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada*
[b]*Department of Medicine, McMaster University, Hamilton, Ontario, Canada*
[c]*MAGIC Evidence Ecosystem Foundation, Oslo, Norway*
[d]*Faculty of Dentistry, Department of Pediatric Dentistry and Orthodontics, Universidad de La Frontera, Temuco, Chile*
[e]*Department of Anesthesia, McMaster University, Hamilton, Ontario, Canada*
[f]*Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada*
[g]*Michael G. DeGroote National Pain Center, McMaster University, Ontario, Canada*
[h]*Department of Biomedical Informatics, and Medical Statistics, Medical Research Institute, Alexandria University, Alexandria, Egypt*
[i]*Chinese Evidence-Based Medicine Center, West China Hospital, Sichuan University, Chengdu, Sichuan, China*

Accepted 26 August 2025; Published online 22 September 2025

Abstract

Objectives: To explore the extent to which trustworthiness of treatment guidelines has changed since the introduction of Institute of Medicine (IOM) standards.

Study Design and Setting: In this systematic survey, we searched Medline, Embase, PsycINFO, and Trip from January to December 2010 (pre-IOM) and 2022 (post-IOM) for guidelines that were informed by a systematic review of evidence, written in English, and included recommendations for the treatment of any health conditions in individuals of any age. Using the 10-item modified National Guideline Clearinghouse Extent of Adherence to Trustworthy Standards (NEATS) instrument, paired reviewers independently assessed trustworthiness of 70 randomly selected guidelines from 2010 and 70 from 2022. We calculated mean difference (MD) and corresponding 95% confidence interval (CI) comparing guidelines from 2010 with 2022 for mean score across four critical items, mean score across all items, and score of individual items on 5-point Likert scales in which 1 was least adherence and 5, highest adherence. We also explored change in the proportion of guidelines that achieved a mean score of 4 or higher across critical items.

Results: The average of mean scores increased from 2.28 to 2.74 (MD: 0.46, 95% CI: 0.13−0.79) across critical items and from 2.20 to 2.52 (MD: 0.32, 95% CI: 0.08−0.56) across all items. The proportion of guidelines with a mean score of 4 or higher across critical items increased from 2.9% to 18.6% (absolute difference: 15.7%, 95% CI: 5.8%−25.6%). The items with the greatest increase in the average score included disclosure of funding sources and disclosure and management of financial conflict of interests, followed by a smaller increase in synthesis of evidence, study selection, and search strategy. Other items showed little to no change from 2010.

Conclusion: As assessed by the modified NEATS instrument, although trustworthiness of treatment guidelines has improved after the introduction of IOM standards, improvements proved modest and post-IOM guidelines trustworthiness remains suboptimal.

Keywords: Clinical practice guideline; Treatment; Trustworthiness; Methodological quality; IOM standards; Modified NEATS instrument

Funding: This project was supported by the Einstein Foundation Berlin as part of the Einstein Foundation Award for Promoting Quality in Research. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement by, the Einstein Foundation or the award jury.

* Corresponding author. Department of Health Research Methods, Evidence, and Impact, McMaster University, 1280 Main St W, Hamilton, Ontario L8S 4L8, Canada.
E-mail address: ghadimim@mcmaster.ca (M. Ghadimi).

https://doi.org/10.1016/j.jclinepi.2025.111962

What is new?

Key findings

- This systematic survey found that despite a modest improvement in trustworthiness of treatment guidelines since the introduction of the Institute of Medicine (IOM) standards, post-IOM guidelines trustworthiness remains suboptimal.

What this adds to what is known?

- Evaluations of guidelines from pre-IOM standards reported most had major limitations, particularly in addressing conflict of interests, convening a multidisciplinary guideline panel, using a systematic review of evidence to inform recommendations, and rating the strength of recommendations.
- Studies on the quality of post-IOM guidelines have reported mixed results, and no study has systematically addressed the extent of improved trustworthiness—if any—across a representative group of guidelines.
- This is the first systematic survey that across a representative group of guidelines addressed the extent to which trustworthiness of treatment guidelines has changed since the introduction of IOM standards.

What is the implication and what should change now?

- Findings from this study identify current methodological limitations in the guideline development process and may inform guideline developers regarding the steps that should be taken to improve the rigor of their process.

1. Introduction

Clinical practice guidelines are intended to aid decision-making and improve patient-important outcomes [1]. For optimal usefulness, the development of guidelines should implement rigorous methods, adhering to quality standards. In 2011, the Institute of Medicine (IOM) published standards for the development of trustworthy guidelines [1]. Based on IOM standards [1] a trustworthy guideline has an explicit disclosure of funding sources; is free from biases due to conflict of interests (COIs) of members of the guideline development group (GDG); is developed by a guideline panel comprising relevant clinicians, methodological experts, and populations expected to be affected by the guideline; is informed by a systematic review of evidence that meets IOM standards for Systematic Reviews of Comparative Effectiveness Research [2]; provides an appropriate rating of the strength of recommendations; includes recommendations articulated in a clear standardized form; undergoes external review by relevant stakeholders; and presents adequate plans for updating.

Evaluations of guidelines from pre-IOM standards reported most had major limitations [3–7]. Fewer than half of the guidelines published between 2006 and 2011 met more than 50% of IOM standards with no apparent improvement over time [6]. The trustworthiness domains that suffered most frequently included addressing COIs, convening a multidisciplinary guideline panel, using a systematic review of evidence, and rating the strength of recommendations [3–7].

Since the introduction of IOM standards [1], over 10,000 guidelines have been released [8]. Studies on the quality of post-IOM guidelines have reported mixed results [9–11], and no study has systematically addressed the extent of improved trustworthiness—if any—across a representative group of guidelines. In this study, we therefore addressed the extent to which trustworthiness of guidelines, as reflected in adherence to IOM standards, has changed since their introduction.

2. Methods

We conducted a systematic survey addressing trustworthiness of treatment guidelines released before vs after the introduction of IOM standards. We adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses that are relevant to this systematic survey [12].

2.1. Eligibility criteria

Eligible guidelines were written in English and included de novo or updated recommendations for pharmacologic or nonpharmacologic treatment of any health conditions in individuals of any age that were informed by a systematic review of evidence. To be identified as informed by a systematic review of evidence, guidelines met at least one of the following criteria: (i) explicitly stated that a systematic review of evidence informed the guideline; (ii) presented a list of databases searched and search terms used; (iii) described a study selection process that included number of studies identified through the search and number of studies included in the review; (iv) provided a summary of inclusion and exclusion criteria; or (v) in the guideline development section, cited a methodology guidance paper that described conduct of a systematic review of evidence for developing guidelines. We excluded adaptation of existing guidelines.

2.2. Literature search

We searched Medline, Embase, and PsycINFO from January 2010 to December 2010 (pre-IOM), the year before

IOM standards were published [1], and from January 2022 to December 2022 (post-IOM). An experienced health sciences librarian refined our search strategy for each database (Supplementary 1). To identify guidelines that were not published in medical journals (ie, only available through the guidelines developers' websites), we searched Trip database.

2.3. Sampling and guideline selection

Pairs of trained reviewers independently screened titles and abstracts, followed by full texts from the randomly ordered guidelines within each stratum of journal and non-journal guidelines, with disagreement adjudicated by a third reviewer (M.G.). We continued selection until we achieved the target sample size of 140 eligible guidelines which was determined based on feasibility considerations (70 from 2010 and 70 from 2022). Based on the availability of non-journal guidelines in 2010, 79% ($n = 55$) of guidelines in each sample were published in journals and 21% ($n = 15$) were from developers' websites.

2.4. Data extraction

Pairs of reviewers independently extracted data from each eligible guideline, including country of origin, developing organization, population addressed, clinical condition (based on the International Classification of Diseases 11th Revision [13]), and sources of funding. Reviewers resolved discrepancies by consensus and, when necessary, through discussion by a third reviewer (M.G.).

2.5. Assessment of trustworthiness of guidelines

We defined trustworthiness as the extent to which the guideline development process adhered to the methodological standards proposed by the IOM [1]. To improve the performance of the National Guideline Clearinghouse Extent of Adherence to Trustworthy Standards (NEATS) instrument [14], initially developed on the basis of IOM standards [1], we developed modifications of the instrument [15]. The modified NEATS instrument consists of ten items addressing disclosure of funding sources, disclosure and management of COIs, composition of the GDG, systematic review process (search strategy, study selection, and synthesis of evidence), rating the strength of recommendations, articulation of recommendations, external review process, and plans for updating [15]. The item on disclosure of funding sources was rated using two response options (yes/no), and the item on the composition of the GDG was rated using three response options (yes/no/unknown). All other items were scored on a 5-point Likert scale (1 = lowest adherence and 5 = highest adherence to the item).

Paired reviewers, following an online training session and calibration exercises, independently assessed trustworthiness of each eligible guideline using the modified NEATS instrument [15]. Reviewers recorded their responses to the prompting questions for each item and corresponding score in pilot-tested extraction forms and resolved disagreements by consensus or in discussion with a third reviewer (M.G.).

We evaluated the first treatment recommendation from each guideline in our assessment of relevant items, except for the "rating the strength of recommendations" item, for which we evaluated the first five strong treatment recommendations. There exist different classification systems for the strength of recommendations. For instance, the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach categorizes the strength of recommendation as strong or weak [16], whereas the American College of Cardiology and the American Heart Association categorize the strength of recommendation as strong, moderate, and weak [17]. To ensure consistent evaluation across guidelines, we considered a recommendation as strong if the strength of the recommendation was stronger than weak.

2.6. Outcomes of interest

We chose as the primary outcome change in mean score across critical items from 2010 to 2022. Critical items included disclosure and management of COIs, study selection, synthesis of evidence, and rating the strength of recommendations [15]. Secondary outcomes included change in mean score across all items and change in the proportion of guidelines that achieved a mean score of 4 or higher across critical items from 2010 to 2022. We classified guidelines with a mean score of 4 or higher across critical items as highly trustworthy. A secondary outcome at the individual item level was change in the score of each item from 2010 to 2022. The score of each item ranged from 1 to 5; for items addressing disclosure of funding sources and GDG composition, we scored them as 1 if not met (rated "no" or "unknown") and 5 if met (rated "yes").

2.7. Statistical analysis

We summarized the distribution of continuous variables using mean and SD and categorical variables as frequency and percentage. Using independent samples t-tests with equal variances, we calculated mean difference (MD) and corresponding 95% confidence interval (CI) comparing guidelines published in 2010 vs those from 2022 for mean score across critical items, mean score across all items, and score of individual items. Using multivariable linear regression models, we explored the following prespecified subgroups for the primary outcome: (i) guidelines focused on adults vs pediatrics may have greater improvement in trustworthiness, and (ii) guidelines developed by governmental agencies or the World Health Organization vs professional societies or other groups may have greater improvement in trustworthiness. We found no evidence of departure from the assumptions of linear regression in the models. We

planned to assess the credibility of any apparent difference in the subgroups (ie, a *P* value for the test of interaction of less than 0.1) using the ICEMAN tool [18]. We used Fisher exact test to compare the proportion of guidelines published in 2010 vs 2022 with a mean score of 4 or higher across critical items, and implemented all analyses in Stata (17.0; StataCorp).

3. Results

Among the total sample of 429 full texts reviewed, we included 70 guidelines for the year 2010 and 70 guidelines for the year 2022. Supplementary Figures 1 and 2 present guideline selection flowcharts. Excluded full texts were often guidelines that did not use a systematic review of evidence (31.7%) (Supplementary Tables 1 and 2). Table 1 presents the characteristics of the included guidelines. The majority of guidelines were multinational (55/140, 39.3%), focused on the adult population (83/140, 59.3%), and targeted neoplasms (27/−140, 19.3%). Professional societies predominantly developed the guidelines (100/140, 71.4%). Direct funding from pharmaceutical industries supported the development process of 13 (9.3%) guidelines. Supplementary Tables 3 and 4 present an assessment of trustworthiness for each individual guideline with a summary of the assessments in Supplementary Figure 3.

Table 1. Characteristics of included clinical practice guidelines

Characteristics	Year 2010 (*n* = 70)	Year 2022 (*n* = 70)
Region of origin, n (%)		
Multinational	25 (35.7)	30 (42.9)
Europe	18 (25.7)	18 (25.7)
North America	21 (30.0)	14 (20.0)
South America	0	1 (1.4)
South-East Asia	0	1 (1.4)
Western Pacific	6 (8.6)	6 (8.6)
Target population, n (%)		
Adult	43 (61.4)	40 (57.1)
Pediatrics	6 (8.6)	6 (8.6)
Both	21 (30.0)	24 (34.3)
Developing organization, n (%)		
Professional society	49 (70.0)	51 (72.9)
Governmental agency and WHO	15 (21.4)	11 (15.7)
Other groups[a]	6 (8.6)	8 (11.4)
Funding source, n (%)		
Direct industry funding	9 (12.9)	4 (5.7)
Other sources	32 (45.7)	52 (74.3)
Not reported	29 (41.4)	14 (20.0)
Clinical condition, n (%)		
Blood diseases	1 (1.4)	0
Circulatory system diseases	8 (11.4)	8 (11.4)
Digestive system diseases	5 (7.1)	5 (7.1)
Endocrine, nutritional or metabolic diseases	2 (2.9)	3 (4.3)
Genitourinary diseases	2 (2.9)	5 (7.1)
Immune system diseases	3 (4.3)	2 (2.9)
Infectious diseases	14 (20)	7 (10.0)
Mental or neurodevelopmental diseases	1 (1.4)	2 (2.9)
Musculoskeletal diseases	4 (5.7)	6 (8.6)
Neoplasms	14 (20)	13 (18.6)
Nervous system diseases	7 (10)	7 (10.0)
Neonatal diseases	0 (0)	1 (1.4)
Pain management	2 (2.9)	2 (2.9)
Pregnancy, prenatal conditions, and childbirth	2 (2.9)	2 (2.9)
Respiratory diseases	1 (1.4)	1 (1.4)
Skin diseases	2 (2.9)	4 (5.7)
Sleep-wake diseases	2 (2.9)	1 (1.4)
Visual diseases	0	1 (1.4)

WHO, World Health Organization.
[a] Other groups that developed guidelines included study groups, non-for-profit or charity organizations.

3.1. Change in mean score across critical items

Figure 1A displays the distribution of mean scores across critical items. The average of mean scores across critical items was 2.28 in 2010 and increased to 2.74 in 2022 (MD: 0.46, 95% CI: 0.13−0.79) (Table 2).

3.2. Change in mean score across all items

Figure 1B presents the distribution of mean scores across all items. The average of mean scores across all items was 2.20 in 2010 and increased to 2.52 in 2022 (MD: 0.32, 95% CI: 0.08−0.56) (Table 2).

3.3. Change in the proportion of guidelines with high trustworthiness

The proportion of guidelines with a mean score of 4 or higher across critical items was 2.9% in 2010 and increased to 18.6% in 2022 (absolute difference: 15.7%, 95% CI: 5.8%−25.6%) (Table 2).

3.4. Change in score at the individual item level

The items with the greatest increase in the average score from 2010 to 2022 included disclosure of funding sources (MD: 0.86, 95% CI: 0.25−1.46) and disclosure and management of financial COIs (MD: 0.84, 95% CI: 0.20−1.48), followed by smaller increase in synthesis of evidence (MD: 0.53, 95% CI: 0.10−0.96), study selection (MD: 0.49, 95% CI: 0.04−0.93), and search strategy (MD: 0.36, 95% CI: −0.09 to 0.81). Other items showed little to no change from 2010 (Table 2). Supplementary Table 5 summarizes the limitations of guidelines in adherence to each item.

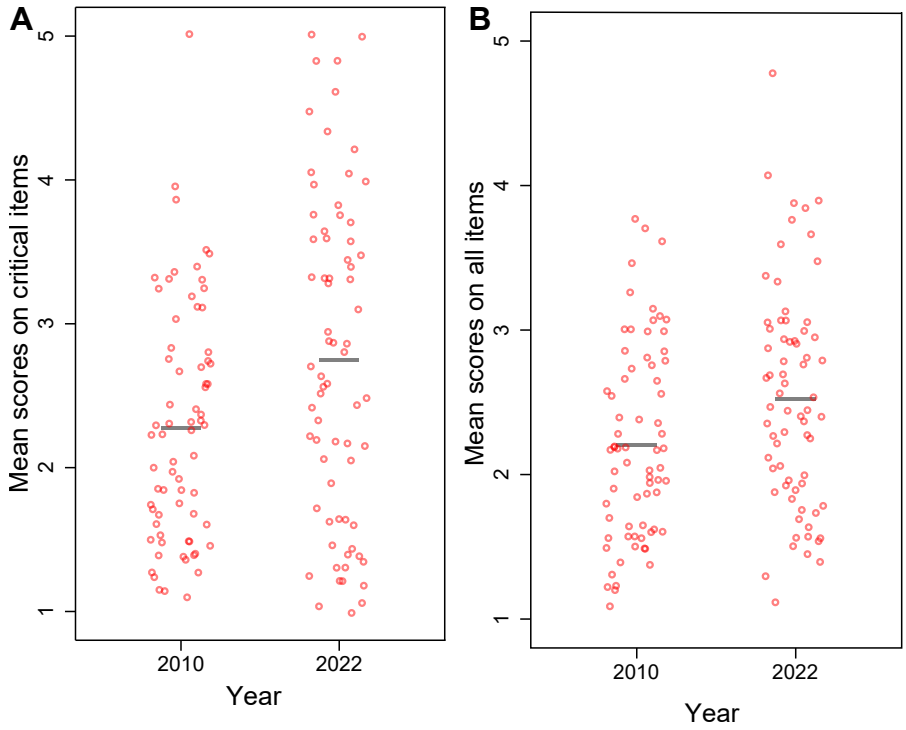

Figure. Distribution of mean scores (A) across critical items and (B) across all items of the modified NEATS instrument. The reference horizontal lines represent the median of mean scores for each year. Critical items included disclosure and management of conflict of interests, study selection, synthesis of evidence, and rating the strength of recommendations. NEATS, National Guideline Clearinghouse Extent of Adherence to Trustworthy Standards.

3.5. Other analyses

The subgroup analysis showed no apparent differences in the magnitude of change in mean score across critical items based on the target population or developers of the guidelines (Supplementary Fig 4).

4. Discussion

4.1. Principal findings

This systematic survey found that overall adherence of treatment guidelines to the trustworthiness standards, as assessed by the mean scores across critical items and all items of the modified NEATS instrument [15], has improved after the introduction of the IOM criteria [1]. The improvement is, however, small and reflects moderate overall adherence to the trustworthiness standards among post-IOM guidelines. Although the proportion of treatment guidelines demonstrating high trustworthiness has increased since IOM standards [1], fewer than 20% of post-IOM guidelines are of high trustworthiness.

Among the domains of trustworthiness, the greatest improvement occurred in adherence to disclosure of funding sources and disclosure and management of financial COIs. Improvement in the COI domain mainly represents increased transparency in disclosing COIs rather than more appropriate strategies to handle them. Adherence to the systematic review domain showed smaller improvement. Limited improvement in the search strategy item reflects the challenge of keeping pace with the rapid growth of evidence, which might have hindered timely updates of the search. Within the study selection process, improvement occurred in defining the eligibility criteria, whereas duplicate screening of retrieved studies decreased from pre-IOM, which might be due to time constraints and limited availability of skilled reviewers amid an increased production of guidelines after IOM. Improvement in the synthesis of evidence item mainly comes from a more appropriate rating of the certainty of evidence among post-IOM guidelines. Other domains of guideline trustworthiness proved little to no change from pre-IOM.

4.2. Strengths and limitations

This study has several strengths. First, to ensure a representative sample of guidelines, we included treatment guidelines published in journals as well as guidelines that were only available on the developers' websites. Second, to minimize bias, we performed selection of guidelines and assessment of trustworthiness in duplicate. Third, to address improvement in the overall adherence of guidelines to the trustworthiness standards, as the primary outcome, we focused on the domains with the most important role in developing trustworthy guidelines as judged by a

Table 2. Primary and secondary outcomes

Outcomes of interest	Year 2010 (Pre-IOM) (n = 70)	Year 2022 (Post-IOM) (n = 70)	Absolute difference (95% CI)[b]	*P* value
Primary outcome				
Mean scores across critical items[a], mean (SD)	2.28 (0.81)	2.74 (1.14)	0.46 (0.13–0.79)	0.006
Secondary outcomes				
Mean scores across all items, mean (SD)	2.20 (0.66)	2.52 (0.76)	0.32 (0.08–0.56)	0.008
Mean score of 4 or higher across critical items[a], n (%)	2 (2.9)	13 (18.6)	15.7 (5.8–25.6)	0.005
Scores at item level				
Disclosure of funding sources item, mean (SD)	3.34 (1.98)	4.20 (1.61)	0.86 (0.25–1.46)	0.006
Disclosure and management of COIs item, mean (SD)	2.59 (1.94)	3.43 (1.88)	0.84 (0.20–1.48)	0.010
Composition of the GDG item, mean (SD)	1.34 (1.13)	1.57 (1.41)	0.23 (−0.20 to 0.66)	0.291
Search strategy item, mean (SD)	2.69 (1.37)	3.04 (1.31)	0.36 (−0.09 to 0.81)	0.117
Study selection item, mean (SD)	1.66 (1.20)	2.14 (1.43)	0.49 (0.04–0.93)	0.031
Synthesis of evidence item, mean (SD)	2.24 (1.04)	2.77 (1.49)	0.53 (0.10–0.96)	0.016
Rating the strength of recommendations item, mean (SD)	2.61 (1.53)	2.61 (1.57)	0 (−0.52 to 0.52)	1.000
Articulation of recommendations item, mean (SD)	1.73 (1.14)	1.80 (1.41)	0.07 (−0.36 to 0.50)	0.742
External review item, mean (SD)	2.09 (1.15)	2.10 (1.14)	0.01 (−0.37 to 0.40)	0.941
Updating plan item, mean (SD)	1.74 (0.85)	1.57 (0.81)	−0.17 (−0.45 to 0.11)	0.222

COI, conflict of interest; GDG, guideline development group; IOM, Institute of Medicine.
The range for continuous outcomes was 1 (lowest adherence) to 5 (highest adherence).
[a] Critical items included disclosure and management of COIs, study selection, synthesis of evidence, and rating the strength of recommendations.
[b] We calculated mean difference and corresponding 95% CI for continuous outcomes, and for proportion of guidelines with a mean score of 4 or higher across critical items, difference in the crude proportions with corresponding 95% CI.

consensus of team members with expertise in the guideline methodology [15].

This study also has limitations. First, for feasibility purposes, we sampled only guidelines from two time points, before (2010) and after IOM (2022). Given the possibility for year-to-year variations in trustworthiness of guidelines and potential impacts of coronavirus disease 2019 on guideline development, the observed improvement in trustworthiness may not be generalizable to all post-IOM years. Second, our judgment on critical items was subjective without a formal consensus process. We chose items that, in our view, if badly handled, are most likely to decrease trustworthiness. Third, we only assessed financial COIs and did not address intellectual COIs that can be a source of bias in the development of guidelines [1]. Intellectual COIs are prevalent [19–21] and have been increasingly recognized in COI policies [1,22–25] but remain less frequently addressed than financial COIs in the guideline development process [19,20]. Therefore, it is possible that our assessment overrated guidelines in both years, likely more substantially in 2010, when addressing the ''disclosure and management of COIs'' item. Consequently, the magnitude of improvement in adherence to this item from 2010 may have been underestimated in our study. Fourth, although we modified the NEATS instrument to capture trustworthiness of the guideline development process [15], limitations in reporting may have impacted our assessments for ''disclosure and appropriate management of COIs'', ''search strategy'', and ''study selection'' items. For instance, in the assessment of the ''search strategy'' item, the guidelines achieved the lowest adherence if there was no information on search terms. The authors of the guidelines may have used appropriate search terms that they failed to disclose. It is more likely that guidelines published in 2010 than in 2022 suffer from limited reporting (Supplementary Table 5); therefore, we may have overestimated the magnitude of improvement in adherence to these items from 2010. Fourth, we focused only on a group of recommendations and did not assess all the treatment recommendations in the guidelines. Therefore, it is possible that we overestimated scores for the ''rating the strength of recommendations'' and ''articulation of recommendations'' items in both years; however, the magnitude of change in the scores of these items from 2010 would probably remain unchanged.

4.3. Relation to other studies

Two other studies have explored the change in trustworthiness of guidelines after IOM standards [26,27]. These studies focused on guidelines for specific clinical areas; osteoporosis screening [26] and prescription of physical activity [27]. We adopted a broader scope and included guidelines for treatment of any health conditions.

Although these studies evaluated trustworthiness as measured by adherence of guidelines to IOM standards which assumes guidelines are informed by a systematic review of evidence [1], they did not restrict the inclusion criteria to such guidelines. This makes the assessment of domains related to the appropriateness of the systematic review process irrelevant for guidelines that were not informed by a systematic review of evidence. We did not include such guidelines.

The assessment of trustworthiness in these studies [26,27] had important limitations, including failure to consider the relative importance of criteria under each IOM standard [1] in guideline trustworthiness, evaluating the reporting quality rather than trustworthiness for some standards, and vulnerability to subjective assessment. We addressed these limitations by using the rigorously developed modified NEATS instrument that addresses the appropriateness of the process relevant to each IOM standard using hierarchical algorithms with explicit prompting questions [15].

4.4. Implications of our findings

The findings of this study suggest that despite the improvement in trustworthiness of treatment guidelines since the introduction of IOM standards [1], post-IOM guidelines trustworthiness remains limited. Given the limitations of post-IOM guidelines in our study, guideline developers should focus on improving their process in a number of ways. These include first appointing at least one chair without financial COIs to lead the GDG. Second, recruiting a multidisciplinary GDG that includes methodologists and patient partners. Third, improving the rigor of the systematic review of evidence that informs recommendations by performing an up-to-date search with an adequate selection of databases, defining specific and unambiguous eligibility criteria to direct the study selection, screening retrieved studies in duplicate and independently, presenting quantitative or qualitative synthesis of included studies, and rating the certainty of the evidence for each prioritized outcome. Fourth, rating the strength of recommendations using a structured framework of moving from evidence to recommendations that ideally implements the GRADE approach [28] and avoids issuing strong recommendations based on low or very low certainty evidence except in specific scenarios [28]. Finally, submitting draft guidelines for external review and devising plans for updating guidelines as soon as new practice-changing evidence becomes available to create "living" guidelines [29].

In addition, our findings suggest that guideline developers should provide greater transparency with respect to the reporting of their guideline development process. This particularly includes identifying the roles of GDG members, including chair or co-chairs, describing the details of the systematic review process, and providing explicit statements for the GDG's justification on the balance of benefits and harms of intervention vs comparator (s) linked with the certainty of evidence that informed the strength of recommendations. To facilitate the implementation of recommendations in practice, guideline developers should explicitly identify population and comparators, in addition to the intervention of interest, when articulating recommendations.

Guideline publishers and databases of guidelines can play important roles in improving trustworthiness. Publishers could mandate adherence to IOM standards, whereas databases could provide a public judgment on trustworthiness for individual guidelines and differentiate trustworthy from untrustworthy guidelines.

5. Conclusion

As assessed by the modified NEATS instrument, overall adherence of treatment guidelines to the trustworthiness standards and proportion of treatment guidelines with high trustworthiness have shown modest improvement after the introduction of IOM criteria. Despite the observed improvement, trustworthiness of treatment guidelines published after IOM standards remains suboptimal. Achieving uniformly high trustworthiness in practice guidelines will require considerable additional efforts in identifying and addressing existing barriers.

CRediT authorship contribution statement

Maryam Ghadimi: Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Gordon Guyatt:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Carlos Zaror:** Writing – review & editing, Investigation, Data curation. **João Pedro Lima:** Writing – review & editing, Investigation, Data curation. **Sana Gupta:** Writing – review & editing, Investigation, Data curation. **Hamed Movahed:** Writing – review & editing, Investigation, Data curation. **Wimonchat Tangamornsuksan:** Writing – review & editing, Investigation, Data curation. **Gonzalo Bravo-Soto:** Writing – review & editing, Investigation, Data curation. **Sahrish Masood:** Writing – review & editing, Investigation, Data curation. **Huda Gomaa:** Writing – review & editing, Investigation, Data curation. **Yanjiao Shen:** Writing – review & editing, Investigation, Data curation. **Nima Behravan:** Writing – review & editing, Investigation, Data curation. **Rachel Couban:** Writing – review & editing, Resources. **Romina Brignardello-Petersen:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

M.G. reports financial support was provided by Einstein Foundation Berlin. Co-author, G.B.S., received grants from

the Canadian Institutes of Health Research (CIHR) to conduct a mixed-method study about the pandemic. There are no competing interests for any other author.

Acknowledgments

The authors acknowledge Feryal Momenilandi and Ali Shahabi for their help with the guideline selection process.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2025.111962.

Data availability

The dataset of this study is available as supplementary: Treatment guidelines trustworthiness_dataset

References

[1] Steinberg E, Greenfield S, Wolman DM, Mancher M, Graham R. Clinical practice guidelines we can trust. Washington, DC: The National Academies Press; 2011.
[2] Institute of Medicine. Finding what works in health care: standards for systematic reviw. Washington, DC: The National Academies Press; 2011.
[3] Shaneyfelt TM, Mayo-Smith MF, Rothwangl J. Are guidelines following guidelines? The methodological quality of clinical practice guidelines in the peer-reviewed medical literature. JAMA 1999;281:1900−5. https://doi.org/10.1001/jama.281.20.1900.
[4] Grilli R, Magrini N, Penna A, Mura G, Liberati A. Practice guidelines developed by specialty societies: the need for a critical appraisal. Lancet 2000;355:103−6. https://doi.org/10.1016/S0140-6736(99)02171-6.
[5] Hasenfeld R, Shekelle PG. Is the methodological quality of guidelines declining in the US? Comparison of the quality of US Agency for Health Care Policy and Research (AHCPR) guidelines with those published subsequently. Qual Saf Health Care 2003;12:428−34. https://doi.org/10.1136/qhc.12.6.428.
[6] Kung J, Miller RR, Mackowiak PA. Failure of clinical practice guidelines to meet institute of medicine standards: two more decades of little, if any, progress. Arch Intern Med 2012;172:1628−33. https://doi.org/10.1001/2013.jamainternmed.56.
[7] Kryworuchko J, Stacey D, Bai N, Graham ID. Twelve years of clinical practice guideline development, dissemination and evaluation in Canada (1994 to 2005). Implement Sci 2009;4:49. https://doi.org/10.1186/1748-5908-4-49.
[8] Trip medical database. Available at: https://www.tripdatabase.com/. Accessed January 24, 2025.
[9] Sephien A, Dayto DC, Reljic T, Blonski W, Nerella N, Prida X, et al. A systematic critical appraisal of clinical practice guidelines in heart failure using the AGREE II tool. Am J Cardiol 2023;202:192−8. https://doi.org/10.1016/j.amjcard.2023.06.052.
[10] Amer YS, Titi MA, Godah MW, Wahabi HA, Hneiny L, Abouelkheir MM, et al. International alliance and AGREE-Ment of 71 clinical practice guidelines on the management of critical care patients with COVID-19: a living systematic review. J Clin Epidemiol 2022;142:333−70. https://doi.org/10.1016/j.jclinepi.2021.11.010.
[11] Anwer MA, Al-Fahed OB, Arif SI, Amer YS, Titi MA, Al-Rukban MO. Quality assessment of recent evidence-based clinical practice guidelines for management of type 2 diabetes mellitus in adults using the AGREE II instrument. J Eval Clin Pract 2018;24:166−72. https://doi.org/10.1111/jep.12785.
[12] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. https://doi.org/10.1136/bmj.n71.
[13] World Health Organization. International classification of diseases for mortality and morbidity statistics. Eleventh Revision. 2022. Available at: https://icd.who.int/browse/latest-release/mms/en. Accessed January 10, 2025.
[14] Jue JJ, Cunningham S, Lohr K, Shekelle P, Shiffman R, Robbins C, et al. Developing and testing the agency for healthcare research and quality's national guideline clearinghouse extent of adherence to trustworthy standards (NEATS) instrument. Ann Intern Med 2019;170:480−7. https://doi.org/10.7326/M18-2950.
[15] Ghadimi M, Guyatt G, Brignardello-Petersen R. Modifications to the NEATS instrument for more appropriate and reproducible assessment of guidelines trustworthiness. Clin Public Health Guidel 2024;1:e12025. https://doi.org/10.1002/gin2.12025.
[16] Andrews J, Guyatt G, Oxman AD, Alderson P, Dahm P, Falck-Ytter Y, et al. GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. J Clin Epidemiol 2013;66:719−25. https://doi.org/10.1016/j.jclinepi.2012.03.013.
[17] Halperin JL, Levine GN, Al-Khatib SM, Birtcher KK, Bozkurt B, Brindis RG, et al. Further evolution of the ACC/AHA clinical practice guideline recommendation classification system: a report of the American college of cardiology/American Heart Association Task force on clinical practice guidelines. Circulation 2016;133:1426−8. https://doi.org/10.1161/CIR.0000000000000312.
[18] Schandelmaier S, Briel M, Varadhan R, Schmid CH, Devasenapathy N, Hayward RA, et al. Development of the instrument to assess the credibility of effect modification analyses (ICEMAN) in randomized controlled trials and meta-analyses. CMAJ 2020;192:E901−6. https://doi.org/10.1503/cmaj.200077.
[19] Alhazzani W, Lewis K, Jaeschke R, Rochwerg B, Moller MH, Evans L, et al. Conflicts of interest disclosure forms and management in critical care clinical practice guidelines. Intensive Care Med 2018;44:1691−8. https://doi.org/10.1007/s00134-018-5367-6.
[20] Brems JH, Wagner T, Diamant J, Davis AE, Clayton EW. Intellectual conflicts of interest among cardiology and pulmonology clinical practice guidelines. PLoS One 2023;18:e0288349. https://doi.org/10.1371/journal.pone.0288349.
[21] Akl EA, El-Hachem P, Abou-Haidar H, Neumann I, Schunemann HJ, Guyatt GH. Considering intellectual, in addition to financial, conflicts of interest proved important in a clinical practice guideline: a descriptive study. J Clin Epidemiol 2014;67:1222−8. https://doi.org/10.1016/j.jclinepi.2014.05.006.
[22] Schunemann HJ, Al-Ansary LA, Forland F, Kersten S, Komulainen J, Kopp IB, et al. Guidelines international network: principles for disclosure of interests and management of conflicts in guidelines. Ann Intern Med 2015;163:548−53. https://doi.org/10.7326/M14-1885.
[23] Qaseem A, Wilt TJ, Clinical Guidelines Committee of the American College of P, Forciea MA, Kansagara D, Crandall CJ. Disclosure of interests and management of conflicts of interest in clinical guidelines and guidance statements: methods from the clinical guidelines committee of the American college of physicians. Ann Intern Med 2019;171:354−61. https://doi.org/10.7326/M18-3279.
[24] Ngo-Metzger Q, Moyer V, Grossman D, Ebell M, Woo M, Miller T, et al. Conflicts of interest in clinical guidelines: update of US preventive services task force policies and procedures. Am J Prev Med 2018;54:S70−80. https://doi.org/10.1016/j.amepre.2017.06.034.
[25] National Institute for Health and Care Excellence (NICE). Policy on declaring and managing interests for NICE advisory committees.

2022. Available at: https://www.nice.org.uk/Media/Default/About/Who-we-are/Policies-and-procedures/declaration-of-interests-policy.pdf. Accessed February 3, 2025.

[26] Hayawi LM, Graham ID, Tugwell P, Yousef Abdelrazeq S. Screening for osteoporosis: a systematic assessment of the quality and content of clinical practice guidelines, using the AGREE II instrument and the IOM standards for trustworthy guidelines. PLoS One 2018;13:e0208251. https://doi.org/10.1371/journal.pone.0208251.

[27] Siedler MR, Lamadrid P, Humphries MN, Mustafa RA, Falck-Ytter Y, Dahm P, et al. The quality of physical activity guidelines, but not the specificity of their recommendations, has improved over time: a systematic review and critical appraisal. Appl Physiol Nutr Metab 2021;46:34−45. https://doi.org/10.1139/apnm-2020-0378.

[28] Andrews JC, Schunemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, et al. GRADE guidelines: 15. Going from evidence to recommendation-determinants of a recommendation's direction and strength. J Clin Epidemiol 2013;66:726−35. https://doi.org/10.1016/j.jclinepi.2013.02.003.

[29] El Mikati IK, Khabsa J, Harb T, Khamis M, Agarwal A, Pardo-Hernandez H, et al. A framework for the development of living practice guidelines in health care. Ann Intern Med 2022;175:1154−60. https://doi.org/10.7326/M22-0514.