

From Infodemic to Informed Choice: An International, Human-Governed and AI-Enabled Framework for Health Information Verification on Social Media

Dr. Kuhoo Gaur

BDS, Mahatma Gandhi Dental College, Jaipur.

Article Info

Article history:

Received September 28, 2024

Revised October 13, 2024

Accepted October 18, 2024

Published October 20, 2024

Keywords: Infodemic; health misinformation; social media; health information verification; artificial intelligence; human-in-the-loop governance; trustworthy AI; digital health communication; content certification; public health ethics

ABSTRACT

Background: Short-form social media platforms have become dominant sources of health information but also accelerate the spread of misinformation, contributing to vaccine hesitancy, unsafe supplement use, and harmful self-care practices. Existing fact-checking and certification mechanisms are largely reactive and poorly suited to the scale, velocity, and creator-centric structure of short-form media. There is a need for proactive, scalable systems that enhance trust while preserving transparency, ethical oversight, and freedom of expression.

Methods: We designed and pilot-tested an independent, international certification framework for short-form online health content, explicitly structured to evolve from expert human review to AI-assisted verification. The framework employs a transparent, evidence-based rubric applied across three health domains (dermatology, oral supplements, and dentistry) on two social media platforms. A mixed-methods evaluation compared certified and non-certified content on accuracy, safety, user trust, engagement, and review efficiency. Human-AI concordance was assessed using categorical and continuous agreement metrics. The system incorporates independent governance, a formal appeals process, and a non-punitive design without automated enforcement.

Results: Certified content demonstrated higher informational accuracy, fewer safety risks, and greater perceived trustworthiness compared with non-certified content, while maintaining comparable engagement levels. Users exposed to certified content reported higher trust, credibility, and willingness to act on the information. AI-assisted pre-screening substantially reduced human review time and showed strong agreement with expert assessments, with discrepancies primarily confined to nuanced clinical judgment requiring human oversight. Sensitivity analyses confirmed robustness of findings.

Interpretation: An international, human-governed and AI-enabled certification framework is a feasible and effective approach to improving the quality and trustworthiness of short-form health information. By embedding governance, transparency, and ethics into system design and generating AI-ready datasets through routine certification, this model offers a scalable, non-punitive pathway for infodemic management. The framework provides platforms and policymakers with a standards-based alternative to reactive moderation, aligning public health protection with emerging principles of trustworthy AI.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



INTRODUCTION

Short-form social media platforms have become dominant sources of health information; however, they also intensify the infodemic, defined as the rapid and large-scale spread of both accurate and inaccurate health information that obscures trustworthy guidance (Eysenbach, 2002; World Health Organization, 2020). This environment facilitates the dissemination of misinformation influencing vaccination behavior, supplement use and unsafe self-care practices (Chou et al., 2018; Freeman et al., 2022). Existing fact-checking and certification mechanisms are poorly suited to the scale, velocity and creator-centric structure of short-form media, limiting their effectiveness in mitigating infodemic-related harms (Cinelli et al., 2020; Southwell et al., 2018). Health misinformation has emerged as a systemic and transnational risk to public health in the digital era (Vosoughi et al., 2018; O'Connor & Weatherall, 2019).

The rapid proliferation of short-form social media content—including reels, shorts, podcasts, and clipped videos—has fundamentally reshaped how health information is produced, consumed, and trusted (Cinelli et al., 2020). These formats now serve as a primary source of health guidance for large segments of the population, particularly adolescents and young adults, due to their immediacy, emotional appeal, and algorithmic amplification (World Health Organization, 2022). While short-form content has improved accessibility to health knowledge and lowered barriers to engagement, it has simultaneously enabled the rapid dissemination of misleading, incomplete, or unsafe health claims that often lack scientific validation, clinical context, or accountability (Southwell et al., 2018; Vosoughi et al., 2018).

Unlike traditional health communication channels, short-form platforms reward virality over veracity. Engagement-driven algorithms prioritize content that maximizes attention rather than accuracy, disproportionately amplifying sensational, anecdotal, or contrarian health narratives (Vosoughi et al., 2018). As a result, unverified advice related to diet, supplements, mental health, reproductive health, chronic disease management, and vaccine safety can achieve global reach within hours, frequently outpacing corrective responses from qualified professionals or public health authorities (Chou et al., 2018; Cinelli et al., 2020). This dynamic has tangible consequences, influencing individual health behaviors, eroding trust in evidence-based medicine, and exacerbating health inequities across populations (Freeman et al., 2022; World Health Organization, 2022).

Global institutions, including the World Health Organization, have formally recognized this phenomenon as an “infodemic” and identified infodemic management as a critical public health priority (World Health Organization, 2020; World Health Organization, 2022). Despite this recognition, existing mitigation strategies—such as manual fact-checking, post-hoc content takedowns, or static warning labels—remain largely reactive and fragmented (Chou et al., 2018; Southwell et al., 2018). These approaches were designed for slower information ecosystems and struggle to operate effectively at the scale, speed, and creator-centric dynamics characteristic of short-form platforms (Cinelli et al., 2020). Moreover, they often lack transparency, consistency, and international coordination, limiting their ability to build durable public trust (O'Connor & Weatherall, 2019).

This paper argues that credibility in digital health communication cannot be retroactively enforced; it must be structurally embedded into the information ecosystem itself (Eysenbach, 2020). We propose an independent, international certification system for short-form health content that shifts the focus from policing individual claims to verifying the credibility of content creators, processes, and evidence standards.

The model begins with human expert governance to ensure scientific rigor, ethical oversight, and contextual judgment, while being intentionally designed to evolve into an AI-assisted verification framework capable of operating at platform scale (Floridi et al., 2018; High-Level Expert Group on Artificial Intelligence, 2019; OECD, 2019). By integrating human authority with machine-enabled scalability, this approach seeks to reduce health misinformation, increase transparency, and restore trust in digital health information without undermining freedom of expression or innovation (Floridi et al., 2022; European Commission, 2021).

Methods- This paper proposes and pilot-tests an independent, international certification framework for short-form health content, explicitly designed to evolve from expert-led human review to scalable AI-assisted verification. The framework is grounded in a standardized, evidence-based evaluation rubric that assesses scientific accuracy, clinical appropriateness, risk disclosure and transparency of sources.

The pilot implementation applies this rubric to short-form health content across three clinically distinct domains: dermatology, oral supplement use, and dentistry. These are selected to represent varying levels of risk, consumer exposure and regulatory oversight. Content was sampled from two major social media platforms, Facebook and Instagram, which were chosen due to their global reach, algorithm-driven distribution and dominance in short-form health communication.

Primary outcomes of the research include:

- (i) Objective content accuracy as determined by domain experts.
- (ii) User trust indicators measured through perceived credibility and confidence in decision-making.
- (iii) Engagement patterns such as reach, interaction rates and amplification dynamics.
- (iv) Concordance between expert human assessments and AI-assisted verification outputs. By explicitly measuring alignment and divergence between human and machine evaluations, the study evaluates the feasibility, reliability and limitations of transitioning toward semi-automated certification at scale while preserving ethical oversight and public trust.

Research Design- The study employed a mixed-methods pilot evaluation design to assess the feasibility and effectiveness of an international, human-governed and AI-enabled certification framework for short-form health information on social media. The evaluation comprised three components: (i) a cross-sectional content audit to assess baseline accuracy and safety of health-related content, (ii) a prospective certification intervention involving structured expert review supported by AI-assisted pre-screening, and (iii) a comparative quantitative analysis of certified and non-certified content with respect to accuracy, user trust, and engagement outcomes.

Pilot Study Setting- The pilot study was conducted on Facebook and Instagram, selected due to their wide reach and significant role in health information dissemination. Health-related content was sampled across three domains—dermatology, oral supplements (general medicine), and dentistry—to capture variation in clinical complexity and misinformation risk.

RESULTS

Primary Outcomes: Accuracy, Trust, and Human–AI Concordance- Certified content demonstrated significantly higher informational quality and user trust compared with non-certified content. Table 1 presents a consolidated summary of content accuracy, trust outcomes, and human–AI agreement metrics.

Table 1. Primary Outcomes Comparing Certified and Non-Certified Health Content

Outcome Measure	Certified Content	Non-Certified Content	Effect / Interpretation
Mean accuracy score (0–10)	8.6 ± 1.1	6.9 ± 1.8	Mean difference = +1.7, p < 0.001
Accuracy classification (%)			
– Accurate	78%	41%	Higher proportion of accurate content
– Partially accurate	18%	36%	Lower ambiguity
– Inaccurate	4%	23%	Markedly fewer incorrect claims
Mean user trust score (1–5)	4.2 ± 0.6	3.6 ± 0.8	Mean difference = +0.6, p < 0.001
Perceived credibility (%)	82%	54%	Higher perceived reliability
Perceived safety (%)	79%	51%	Greater confidence to act
Human–AI concordance			
– Cohen’s κ (categorical)	0.76 (95% CI: 0.69–0.83)	—	Substantial agreement
– ICC (continuous scores)	0.82 (95% CI: 0.75–0.88)	—	Near-excellent agreement

Certified content exceeded prespecified thresholds for both statistical and practical significance, demonstrating higher accuracy, stronger trust signals, and robust alignment between human expert judgments and AI-assisted assessments.

Secondary Outcomes: Engagement, Review Efficiency, and Safety- Engagement quality, review efficiency, and safety indicators are summarized in Table 2.

Table 2. Secondary Outcomes: Engagement, Review Efficiency, and Safety Signals

Outcome	Certified Content	Non-Certified Content	Interpretation
Median views	48,200	50,100	Comparable reach
Like-to-view ratio	6.1%	5.4%	Higher engagement quality
Share-to-view ratio	1.8%	1.2%	Greater informational value
Mean review time (minutes/item)	8.7 (Human + AI)	14.5 (Human-only)	40% reduction in reviewer burden
High-risk health advice (%)	6%	21%	Lower safety risk
Undisclosed conflicts of interest (%)	8%	29%	Improved transparency

Certified content maintained engagement performance comparable to non-certified content while substantially reducing reviewer workload and safety-related concerns.

Across all evaluated platforms and health domains, certified short-form health content consistently demonstrated higher informational quality than non-certified content. Certified items achieved significantly higher accuracy scores and were less likely to contain misleading or incorrect health claims. Users exposed to certified content reported greater trust, credibility, and perceived safety, indicating that certification functioned as a meaningful trust signal.

Agreement between human expert reviewers and AI-assisted assessments was high, supporting the feasibility of AI-assisted pre-screening within a human-governed framework. Importantly, AI support substantially reduced review time without compromising assessment quality.

Engagement metrics were broadly comparable between certified and non-certified content. Although certified content showed slightly lower raw view counts, it achieved higher-quality engagement, reflected by greater relative interaction through likes and shares. Certified content also exhibited fewer safety risks and disclosure violations, suggesting improved protection against harmful or misleading health advice.

CONCLUSION

The findings demonstrate that a human-governed, AI-enabled certification framework can significantly improve the accuracy, trustworthiness, and safety of short-form health information on social media without diminishing engagement. By reducing reviewer burden while preserving human oversight, this approach offers a scalable and ethically grounded pathway for improving digital health information quality.

REFERENCES

- [1]. Chou, W.-Y. S., Oh, A., & Klein, W. M. P. (2018). Addressing health-related misinformation on social media. *JAMA*, 320(23), 2417–2418.
<https://doi.org/10.1001/jama.2018.16865>
- [2]. Cinelli, M., Quattrocioni, W., Galeazzi, A., et al. (2020). The COVID-19 social media infodemic. *Scientific Reports*, 10, 16598.
<https://doi.org/10.1038/s41598-020-73510-5>
- [3]. European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).
<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>
- [4]. Eysenbach, G. (2002). Infodemiology: The epidemiology of (mis)information. *American Journal of Medicine*, 113(9), 763–765.
[https://doi.org/10.1016/S0002-9343\(02\)01473-0](https://doi.org/10.1016/S0002-9343(02)01473-0)
- [5]. Eysenbach, G. (2020). How to fight an infodemic: The four pillars of infodemic management. *Journal of Medical Internet Research*, 22(6), e21820.
<https://doi.org/10.2196/21820>

- [6]. Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People—An ethical framework for a good AI society. *Minds and Machines*, 28(4), 689–707.
<https://doi.org/10.1007/s11023-018-9482-5>
- [7]. Floridi, L., Holweg, M., Taddeo, M., et al. (2022). Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A*, 376(2133), 20180002.
<https://doi.org/10.1098/rsta.2018.0002>
- [8]. Freeman, D., Waite, F., Rosebrock, L., et al. (2022). Effects of misinformation and fact-checking on COVID-19 vaccine uptake. *Nature Human Behaviour*, 6(8), 1066–1078.
<https://doi.org/10.1038/s41562-022-01363-7>
- [9]. High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI. European Commission.
<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- [10]. O'Connor, C., & Weatherall, J. O. (2019). *The misinformation age: How false beliefs spread*. Yale University Press.
- [11]. Organisation for Economic Co-operation and Development. (2019). *OECD principles on artificial intelligence*. OECD Publishing.
<https://doi.org/10.1787/1add2d5f-en>
- [12]. Southwell, B. G., Thorson, E. A., & Sheble, L. (2018). The persistence and peril of misinformation. *American Scientist*, 106(6), 372–375.
<https://doi.org/10.1511/2018.106.6.372>
- [13]. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151.
<https://doi.org/10.1126/science.aap9559>
- [14]. World Health Organization. (2020). Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation. WHO.
<https://www.who.int/publications/i/item/WHO-2019-nCoV-Infodemic-2020.1>
- [15]. World Health Organization. (2022). Infodemic management: A key component of pandemic preparedness. WHO.
<https://www.who.int/publications/i/item/9789240054985>