British Journal of Hospital Medicine

*Br. J. Hosp. Med. (Lond)* **2026**; 87(9): 51991
https://doi.org/10.31083/BJHM51991

*Review*

# Artificial Intelligence and Health Misinformation in Healthcare: Opportunities, Limitations and Future Directions

Hasan H Alsararatee[1,2,*], Ahmad Abu Allam[3]

[1]Acute Medicine, Northampton General Hospital NHS Trust, NN1 5BD Northampton, UK
[2]Advanced Practice Faculty, Buckinghamshire New University, HP11 2JZ High Wycombe, UK
[3]School of the Environment and Life Sciences, Faculty of Science and Health, University of Portsmouth, PO1 2UP Portsmouth, UK
*Correspondence: hasan.alsararatee@nhs.net (Hasan H Alsararatee)

Academic Editor: Tonny Veenith
Submitted: 17 March 2026 Revised: 7 August 2026 Accepted: 20 August 2026 Published: 18 September 2026

**Abstract**

The rapid digitalisation of health information has transformed how individuals access medical knowledge but has also facilitated the widespread dissemination of health misinformation. The World Health Organization describes this phenomenon as an infodemic, characterised by an overabundance of information that includes both accurate and misleading content, making it difficult for individuals to identify reliable sources of health advice. Exposure to misinformation has been linked to vaccine hesitancy, reduced adherence to public health guidance and increased engagement with unverified treatments. Artificial intelligence (AI) technologies are increasingly being explored as tools to detect, monitor and counter misleading health narratives across digital platforms. This article critically explores current evidence on AI-based approaches to health misinformation governance, including machine learning, natural language processing, deep learning, large language models, multimodal systems and social media surveillance. Evidence suggests that AI can help identify misleading narratives and emerging public concerns, but current studies rarely distinguish intentional disinformation from unintentional misinformation when reporting model performance. Significant challenges include dataset bias, restricted generalisability across linguistic and cultural contexts, platform-specific model development, limited external validation, privacy concerns, uncertain clinical relevance and amplification through engagement-driven recommendation systems. The article argues that AI should be understood as one component within an iterative, risk-stratified governance framework that combines healthcare expertise, public health interpretation, transparent oversight, platform accountability, tailored digital health literacy interventions and evaluation of real-world public health outcomes.

**Keywords:** artificial intelligence; health communication; misinformation; social media; public health surveillance

## 1. Introduction

The digital transformation of healthcare information has fundamentally changed how individuals access medical knowledge [1]. While digital platforms have increased access to health information, they have also facilitated the rapid dissemination of misinformation, which can undermine evidence-based healthcare practice [2]. The World Health Organization (WHO) describes this phenomenon as an infodemic, characterised by an overabundance of information that includes both accurate and misleading content, making it difficult for individuals to identify reliable sources of health advice [3]. Health misinformation refers to false, misleading or unsupported information relating to health, disease, treatment, prevention or healthcare. It may be shared unintentionally, although its effects can still be harmful. Disinformation refers to false or misleading information shared with deliberate intent to deceive. Throughout this article, health misinformation is used as the umbrella term unless deliberate intent is specifically relevant, because most AI detection studies classify content according to veracity, misleadingness or evidential support rather than intent. This convention is necessary because intent is often not observable from a single post, image or video and may require assessment of source behaviour, coordination, amplification and monetisation.

During the COVID-19 pandemic, misinformation relating to vaccines, treatments and disease origins circulated widely across social media platforms, contributing to vaccine hesitancy and reduced adherence to public health interventions [4,5]. Previous studies have demonstrated that exposure to misinformation can influence public behaviour, including reduced vaccination uptake and increased engagement with unverified treatments [6,7]. However, COVID-19 misinformation should not be treated as fully representative of all health misinformation. Similar concerns arise in relation to cancer, vaccination beyond COVID-19, reproductive health, nutrition, mental health, chronic disease management and alternative treatments. The scale and speed of misinformation dissemination create significant challenges for traditional public health communication. Social media platforms can expand content through algorithmic recommendation systems that prioritise engagement rather than accuracy. These dynamics may create environments in which misinformation spreads rapidly through highly connected online communities [8].

AI technologies have therefore been increasingly explored as tools for monitoring and mitigating health misinformation. Machine learning algorithms can analyse large volumes of digital data, classify potentially misleading information and detect emerging narratives across digital platforms. However, despite rapid technological developments, the integration of AI into public health communication remains complex and raises important methodological, ethical and governance questions.

Within the United Kingdom, the National AI Strategy emphasises the need for responsible AI development and governance across sectors, including healthcare [9]. Within the health system, initiatives such as the NHS AI Lab aim to support the safe integration of AI into clinical practice and healthcare services [10]. At the same time, the Online Safety Act 2023 introduces regulatory responsibilities for digital platforms to address harmful online content [11]. These developments show that technological solutions must be accompanied by governance frameworks that support safe, transparent and ethical implementation. This article critically explores the evidence on AI-based approaches to detecting, monitoring and countering health misinformation. It focuses on machine learning, natural language processing, deep learning, large language models, multimodal models, social media surveillance and human oversight. It also considers practical issues relating to healthcare communication, public health interpretation, privacy, cost, platform accountability and future research.

## 2. Scope and Conceptual Boundaries

This article is a critical narrative review. Its purpose is not to provide a systematic review of all AI misinformation studies, but to synthesise key developments, limitations and governance implications for healthcare and public health communication. The review focuses on AI technologies used to identify, classify, monitor or respond to health misinformation. These include traditional machine learning, natural language processing, transformer-based models, deep learning, multimodal systems, large language models and human-in-the-loop approaches. It also discusses platform recommendation algorithms because these systems shape the visibility and spread of health misinformation. The article focuses on misinformation across digital health environments, including social media platforms, online forums and digital communication systems. Health topics include infectious diseases, vaccination, cancer, chronic disease, reproductive health, mental health, nutrition and alternative treatments. Although many studies emerged from the COVID-19 context, the implications extend beyond pandemic communication.

A conceptual distinction is required between several related terms. Infodemiology refers to the study of the distribution and determinants of health information online and its influence on public health behaviour [7]. Social media surveillance refers to the use of digital data to monitor online health-related conversations and trends. Social listening involves the analysis of public concerns, questions and information needs. Information voids are areas where public demand for reliable health information exceeds the availability of accessible and trustworthy guidance [12]. AI-based detection refers to computational systems used to classify or flag potentially misleading information.

These concepts are connected but should not be treated as interchangeable. Detection identifies potentially misleading content. Social listening identifies emerging concerns and information needs. Infodemiology examines wider patterns in the digital information environment. Governance determines how these signals are interpreted and acted upon.

## 3. Proposed Framework for AI-Enabled Health Misinformation Governance

AI should be understood as one component of a wider health misinformation governance system. A practical framework may include six linked stages, but these stages should be understood as iterative and non-linear rather than as a one-way sequence. Detection identifies potentially misleading claims, narratives, images, videos or patterns of coordinated activity. Validation then requires healthcare professionals, public health experts, subject specialists or trained reviewers to assess clinical meaning and evidential support. Interpretation is needed because online discussion volume does not necessarily represent clinical risk, disease burden or population-level health need. Communication response involves providing timely, clear and accessible information that addresses public concerns. Platform accountability requires attention to ranking systems, recommendation algorithms, advertising models and monetisation. Evaluation should assess not only technical performance but also public health outcomes, equity, trust, cost and safety. In practice, evaluation findings should feed back into detection thresholds, annotated datasets, model retraining, validation rules, communication priorities and platform escalation pathways. Several stages may therefore operate in parallel during an active infodemic response.

For example, if social listening identifies a rise in posts alleging that human papillomavirus vaccination causes infertility, AI tools may first flag the narrative and map its spread across platforms. Public health and vaccination experts would then validate whether the content is false, misleading, uncertain or a genuine unanswered concern. Interpretation would compare online signals with vaccine uptake data, service enquiries and community feedback. Communication teams could then produce targeted, accessible messages that address fertility concerns, while platforms may label, deprioritise or restrict monetisation of demonstrably false claims. Evaluation would assess changes in exposure to the claim, engagement with authoritative information, public understanding, service enquiries and vaccination behaviour. These findings would then refine future keyword

lists, annotated training examples, model thresholds and communication messages, creating a feedback loop rather than a linear workflow [12,13,14,15].

## 4. AI for Detecting Health Misinformation

One of the most widely studied applications of AI in this domain is the automated classification of misinformation using machine learning models. These systems analyse textual, visual, semantic and network-based data to determine whether content may be accurate, misleading or unsupported. Wang et al. [16] developed a multimodal deep learning model for detecting antivaccine misinformation on Instagram, using visual and textual features from images, captions, text embedded in images through optical character recognition, and hashtags. The study used 31,282 Instagram posts collected between January 2016 and October 2019 and reported above 97% testing accuracy across 30 experiments [16]. This supports the value of multimodal approaches for image-rich platforms, although the findings remain platform-specific and focused on antivaccine content rather than health misinformation more broadly. Similarly, Wang et al. [17] proposed a hybrid Bidirectional Encoder Representations from Transformers–Long Short-Term Memory model for misinformation detection in mobile social networks. Their model achieved 93.51% accuracy, 91.96% recall and an F1-score of 92.73%, and a controlled user study involving 100 participants found higher misinformation detection accuracy in the experimental group than in the control group, 89.4% compared with 74.2% [17]. These findings suggest that transformer-based and sequence-learning architectures can support early-stage misinformation detection using textual content, while also showing potential value for digital literacy interventions.

More complex hybrid systems have also been proposed. Zhang et al. [18] developed the Tri-Intelligence framework, which integrates deep neural networks, large language models and crowdsourced human intelligence to detect complex public health misinformation, including conspiracy theories, sarcasm and subtly misleading content. In their evaluation of COVID-19-related social media misinformation, TriIntel outperformed deep neural network, large language model and human-AI collaboration baselines. Compared with the best-performing baseline, StreamCollab, TriIntel achieved improvements of 9.1% in Accuracy, 11.6% in F1-score, 15.9% in Cohen's Kappa and 17.2% in Matthews Correlation Coefficient [18]. These findings suggest that hybrid, human-in-the-loop approaches may improve classification performance for complex misinformation; however, their practical value still depends on external validation, scalability, annotation quality and evaluation in real-world public health communication settings.

Du et al. [19] applied machine learning and deep learning algorithms to 28,121 Reddit posts related to human papillomavirus vaccination. A convolutional neural network achieved the strongest performance, with an area under the receiver operating characteristic curve of 0.7943; 7207 posts were classified as vaccine misinformation, most commonly relating to general safety concerns [19]. This supports the feasibility of automated monitoring of vaccine misinformation in online communities, although the findings remain limited to one platform and one vaccine topic. Zhou et al. [20] developed a multi-feature fusion approach for fake news detection by integrating linguistic, semantic and network-based features. Sharifpoor et al. [21] used machine learning and deep learning algorithms to classify COVID-19 misinformation on Twitter/X. Although these studies show technical promise, several methodological limitations must be considered. Many models rely on datasets derived from specific platforms, such as Instagram, Reddit or Twitter/X. This limits generalisability because each platform has different users, norms, content formats and algorithmic structures. A model trained on short text posts may not perform adequately when applied to images, videos, private messaging platforms or long discussion threads.

A further limitation concerns language and cultural context. Valdez et al. [22] demonstrated that vaccine misinformation patterns differ across English and Spanish language communities. Models trained predominantly on English-language data may therefore perform poorly in multilingual or culturally distinct information environments. This is particularly important because health misinformation often draws on local beliefs, idioms, mistrust, political context and community experience. The quality of annotation is also critical. Health misinformation labels may be assigned by researchers, fact-checkers, crowdworkers or automated rules. Yet the boundary between misinformation, uncertainty, personal experience and legitimate scientific debate can be difficult to define. This is especially relevant in emerging health topics where evidence develops over time.

Many studies also report technical performance using accuracy, precision, recall or F1-score. These metrics are useful but incomplete. Accuracy may be misleading when datasets are imbalanced. False positives may suppress legitimate concerns or scientific debate, whereas false negatives may allow harmful misinformation to spread. Therefore, model evaluation should consider the cost and consequences of different types of error. Di Sotto and Viviani [23] emphasised that many detection systems remain experimental and have not been integrated into real-world health communication infrastructures. Consequently, there is a need for studies that examine whether AI-based detection improves public health outcomes such as vaccination uptake, trust in public health messaging, health information literacy and adherence to evidence-based advice.

Current evidence does not support a reliable quantitative comparison of AI model accuracy for intentional disinformation versus unintentional misinformation. The studies reviewed generally report performance against labels

such as misinformation, fake news, vaccine misinformation or misleading health content, rather than against independently verified authorial intent [16,17,18,19,20,21,22,23]. High accuracy in detecting a misleading claim should therefore not be interpreted as evidence that a model can infer deliberate deception. Unintentional misinformation may be detected through claim content, evidence alignment and semantic contradiction, whereas disinformation may require additional analysis of source credibility, coordinated posting, bot-like amplification, repeated narrative seeding, monetisation and cross-platform behaviour. This distinction is particularly important in high-stakes clinical topics, where misclassification may either suppress legitimate patient concern or allow harmful content to remain visible.

## 5. Comparative Analysis of AI Technical Approaches

Different AI approaches have different strengths, limitations and suitable uses. Traditional machine learning models can be useful where datasets are structured and features are clearly defined. They may be less resource-intensive and more interpretable than deep learning models, but they often require manual feature engineering and may struggle with sarcasm, ambiguity and complex narratives. Natural language processing approaches can support topic modelling, sentiment analysis, claim extraction and narrative mapping. These methods are useful for social listening and infodemiology, but they may miss visual misinformation, coded language or culturally specific meanings.

Deep learning and transformer-based models can analyse complex textual patterns and contextual meaning. They may perform well on large datasets but require substantial computational resources and careful validation. Multimodal deep learning can combine text and image analysis, which is useful for platforms such as Instagram and YouTube. However, multimodal models often require large, annotated datasets and may be less interpretable. Large language models offer potential for summarising trends, extracting claims and supporting public health communication. However, they also create risks, including fabricated content, overconfident responses and the reproduction of misleading narratives. Their outputs require human review, particularly where clinical advice or public health action is involved.

Human-in-the-loop systems are likely to be the most appropriate approach for high-risk health misinformation, but the level of human review should match the clinical risk of the content. Crowd annotators and trained moderators may be sufficient for initial sorting or policy tagging of low-risk claims, but they should not be treated as substitutes for clinical or public health judgement. Claims concerning cancer treatment, medicines, pregnancy, vaccine safety, emergency symptoms or self-harm require review by healthcare professionals, public health specialists or topic experts. The revised comparison in Table 1 therefore distinguishes technical approach, risk tier and reviewer expertise rather than presenting human oversight as a single category.

## 6. Social Media Surveillance and Infodemiology

Beyond automated detection models, AI has been used to analyse large-scale digital data to understand how misinformation spreads within online communities. Infodemiology focuses on analysing digital information flows to identify emerging health narratives and behavioural trends [7]. Studies have demonstrated that social media data can provide valuable insight into public attitudes toward health interventions. Pobiruchin et al. [24] analysed Twitter data during the early stages of the COVID-19 pandemic and identified temporal and geographical variation in the dissemination of COVID-19-related information across Europe. Their findings demonstrated that digital information flows can vary substantially across time and place.

Boatman et al. [13] examined the feasibility of using a commercially available social-listening platform to monitor online misinformation about human papillomavirus vaccination in the United States, including national-level monitoring and state-level analysis in Mississippi and Rhode Island. Their findings suggest that social-listening dashboards may support real-time identification of emerging misinformation narratives and geographically specific patterns. However, the study also highlighted important limitations, including uncertain data quality, incomplete representativeness, dependence on social media data-access agreements, imprecise geographic information and variable accuracy of AI-driven sentiment and emotion classification. Boatman et al. [13] should therefore be interpreted as providing evidence for the potential use of social listening in digital public health surveillance, rather than evidence that misinformation directly reduced vaccination uptake. Predictive modelling studies have also explored the relationship between digital data and public health outcomes. Yan et al. [25] developed a Hidden Markov Model based on social media infodemic signals to predict the number of severe and critical COVID-19 cases during the Wuhan lockdown. Such studies suggest that online information patterns may provide useful public health signals. However, social media data should not be treated as equivalent to clinical testing data, epidemiological surveillance data or true disease burden.

Digital datasets often contain unstructured and unreliable data, including bot activity, spam content and automated posts. Social media users are also not a representative sample of the population. Younger and more digitally engaged groups may be overrepresented, while people with limited internet access, lower digital literacy or language barriers may be underrepresented. Therefore, AI-supported surveillance should be interpreted alongside clinical data, epidemiological surveillance, service use patterns and public health expertise.

**Table 1. Comparison of AI approaches for health misinformation governance**.

| AI approach | Strengths | Limitations | Risk tier and reviewer expertise | Suitable use |
|---|---|---|---|---|
| Traditional machine learning | Lower resource requirement; relatively interpretable where features are explicit. | Requires feature engineering; limited handling of sarcasm, ambiguity and changing narratives. | Low to medium risk. Suitable for screening by trained analysts; expert review needed before action on clinical claims. | Initial classification, topic-specific monitoring and low-risk trend detection. |
| Natural language processing | Supports topic modelling, claim extraction, sentiment analysis and narrative mapping. | Limited for images, videos, coded language and culturally specific meanings. | Low to medium risk. Suitable for public health analysts and communications teams reviewing population concerns. | Social listening, information-need assessment and identification of emerging public concerns. |
| Deep learning | Strong performance with large, labelled datasets and complex textual patterns. | Lower interpretability; requires substantial data, computing resources and external validation. | Medium to high risk. Requires clinical or public health validation where outputs may influence advice, escalation or platform action. | Large-scale classification and prioritisation of content for further review. |
| Transformer-based models | Strong contextual language analysis and capacity to detect subtle semantic patterns. | Requires validation across topics, platforms, languages and populations. | Medium to high risk. Specialist review is required for clinical uncertainty, medicines, cancer, pregnancy or vaccine-safety claims. | Text-based detection, claim extraction and triage of potentially harmful misinformation. |
| Multimodal models | Analyses text, images, hashtags and visual features together. | Requires large annotated datasets; can be difficult to interpret and may perform poorly across platforms. | Medium to high risk. Human review should include health-content expertise and awareness of image context. | Visual misinformation on platforms such as Instagram, TikTok and YouTube. |
| Large language models | Can summarise themes, draft communications and support claim extraction. | Risk of hallucination, overconfident outputs and reproduction of misleading narratives. | High risk if used without oversight. Not appropriate as standalone moderation or public-facing clinical advice for cancer treatment, emergencies, medicines, pregnancy, suicide or vaccine safety. | Assisted review, drafting and synthesis only, with expert checking and audit. |
| Human-in-the-loop: crowd annotators or trained moderators | Adds human judgement at scale and can improve labelling consistency for simple claims. | May lack clinical expertise; quality depends on training, guidance and audit. | Low to medium risk. Appropriate for initial sorting, policy tagging and clear non-clinical misinformation. | Low-risk screening, annotation support and escalation of uncertain cases. |
| Human-in-the-loop: healthcare, public health or topic experts | Combines AI scale with professional judgement and contextual interpretation. | Requires workforce capacity, governance, escalation pathways and accountability. | High risk. Necessary for ambiguous, evolving or clinically consequential misinformation. | Cancer treatment claims, medicines advice, vaccine safety, pregnancy, emergency symptoms, self-harm and public health interventions. |

AI, Artificial intelligence.

## 7. AI in Health Communication

AI can also support health communication strategies aimed at countering misinformation. Digital communication initiatives can reach large audiences, but not all such initiatives are AI tools. Therefore, non-AI examples should be distinguished from AI-enabled systems.

AI-enabled chatbots represent one communication tool. The CoronaAI chatbot, evaluated by Abonizio et al. [26], provided automated responses to misinformation circulating during the COVID-19 pandemic in Brazil. Analysis of chatbot interactions revealed that many users raised misinformation-related queries. This suggests a public demand for accessible, timely and evidence-informed digital health information. Cheng [27] examined chatbot-supported anti-misinformation behaviour in the context of vaccine misinformation. This work suggests that chatbot interactions may encourage users to verify health information and challenge misinformation narratives within online communities. However, chatbot-based interventions depend heavily on trust, transparency, usability and accuracy. If users perceive digital tools as biased, unsafe or unreliable, engagement may decline.

AI-assisted communication should therefore include clear safeguards. Messages should be grounded in evidence, reviewed by relevant experts and written in accessible language. AI tools should not replace healthcare professionals or public health communicators. Rather, they may support drafting, triage, summarisation and identification of common concerns. There is also a risk that AI-generated communication may produce inaccurate or overly confident responses. This is particularly important in health contexts where poor advice may lead to delayed treatment, unsafe self-medication or reduced adherence to evidence-based care. Human oversight remains essential.

## 8. Algorithmic Amplification and the Paradox of AI in Misinformation Ecosystems

While AI is increasingly used to identify and address health misinformation, digital platform algorithms may also contribute to its amplification. Social media systems rely heavily on recommendation engines designed to maximise user engagement. These systems may prioritise content likely to generate interaction, including sensational or emotionally charged misinformation narratives. It is important to distinguish between AI models used for misinformation detection and platform algorithms used for content recommendation. Detection models aim to classify or flag content, while recommendation algorithms rank and promote content. Nevertheless, both are part of algorithmic information systems that shape public exposure to health information.

Röchert et al. [28] analysed information networks on YouTube during the early stages of the COVID-19 pandemic and identified informational homogeneity within user communities. Cinelli et al. [8,29] demonstrated that social media platforms can produce highly polarised information networks in which misinformation spreads rapidly through connected user communities, and later described the echo chamber effect, where repeated exposure to similar content may strengthen belief persistence. These findings suggest that correction after misinformation has spread may be insufficient if platform ranking systems continue to reward engagement, emotional salience or group conformity.

The evidence on modified ranking and public-health-weighted algorithms is still developing, and direct health-specific platform trials remain limited. However, empirical studies indicate that algorithmic visibility can be influenced. Pennycook et al. [30] found that a brief accuracy prompt in the context of COVID-19 misinformation nearly tripled truth discernment in subsequent sharing intentions. Matias [31] showed in a field experiment on Reddit that encouraging fact-checking of unreliable news sources reduced the algorithmic promotion of those articles by as many as 25 rank positions out of 300, sufficient to remove many posts from the front page. These studies do not prove that health-content weighting will reliably break echo chambers, but they support the principle that ranking, accuracy cues and community validation can influence exposure to low-credibility content.

In public health settings, practical mitigation pathways may include increasing the ranking visibility of authoritative health information during periods of concern, reducing the ranking priority of demonstrably false claims, applying contextual labels, restricting monetisation of harmful misinformation and establishing rapid escalation routes for high-risk claims. Evaluation should include measurable outcomes such as exposure to low-credibility health content, reach of trusted public health sources, network diversity, re-sharing of false claims, correction uptake, click-through to authoritative material, trust, health literacy, subgroup equity and downstream indicators such as vaccine enquiries, appointment attendance or help-seeking behaviour. These measures must be transparent, auditable and proportionate so that legitimate patient experience, public concern and scientific debate are not misclassified as harmful content.

## 9. AI and the Identification of Information Gaps in Public Health Communication

A further application of AI in public health is the identification of information voids. Information voids arise when reliable health information is absent, inaccessible or insufficient. When individuals seek answers to unresolved health questions and authoritative information is difficult to find, misleading narratives may fill the gap. Purnat et al. [12] developed a methodology for infodemic signal detection during the COVID-19 pandemic, using digital surveillance tools to analyse social media discussions and identify potential information voids. This approach enables public

health organisations to move from reactive correction toward proactive communication.

The WHO Early AI-Supported Response with Social Listening platform integrates multilingual social media monitoring to detect emerging misinformation signals and information gaps across digital environments [14]. By analysing large volumes of online content in multiple languages, such systems can support timely public health responses. The concept of information voids highlights the importance of responsive communication. Public health messaging often focuses on broadcasting verified information through official channels. However, misinformation frequently emerges in response to unanswered public concerns [12,14]. Effective communication therefore requires continuous monitoring of public discourse and the rapid provision of clear, accessible information addressing emerging questions [32]. Nevertheless, AI-supported identification of information voids has limitations. Social media datasets may not represent the needs of populations with limited digital access, lower literacy or limited trust in health institutions. AI insights should therefore be combined with qualitative research, community engagement, patient experience data and public health expertise.

## 10. Ethical and Governance Challenges

The application of AI in health misinformation detection raises significant ethical and governance concerns. A major issue is algorithmic bias. Machine learning models trained on limited datasets may reproduce existing social biases or perform unequally across languages, demographic groups or cultural contexts [33,34]. Di Sotto and Viviani [23] also highlight the importance of data quality and methodological rigour in health misinformation detection. Privacy concerns arise when analysing large-scale social media datasets. Many studies use publicly available data, but public availability does not remove ethical responsibility. Individuals may not expect their posts to be analysed for public health surveillance or AI model development. Researchers and public health organisations must balance potential public health benefits against the need to protect privacy, dignity and trust [33,34].

Practical safeguards include data minimisation, aggregation, de-identification where possible, secure storage, restricted access, retention limits, transparent governance approvals and clear audit processes. Where feasible, privacy-preserving approaches such as federated learning and differential privacy may support future development by reducing the movement or exposure of personal data. These approaches should be considered as potential governance mechanisms rather than assumed solutions, because they require technical capacity and careful validation.

Transparency and accountability are also central. AI systems used in health misinformation governance should be explainable enough for public health teams to understand their limitations. Organisations should define who is responsible for reviewing AI outputs, escalating uncertain cases and deciding whether communication or platform intervention is required. Automated classification should not be the sole basis for content removal or public health action in high-risk or ambiguous cases. Platform responsibility is another key issue. Röchert et al. [28] demonstrated that recommendation algorithms on platforms such as YouTube can create informational homogeneity. Zenone et al. [35] documented the advertising of alternative cancer treatments on social media platforms, illustrating how misinformation can be monetised through digital advertising systems. These findings suggest that technological detection alone is insufficient. Governance must also address platform incentives, advertising systems and accountability structures.

## 11. Implications for Public Health and Healthcare Systems

AI offers opportunities to strengthen public health responses to health misinformation. AI-driven social listening systems can help organisations monitor digital discussions, identify emerging misinformation narratives and respond to information gaps [12,14]. Machine learning classification models can also support analysis of large volumes of digital data that would be impossible to review manually [16,21]. However, the strongest relevance of these tools is to public health communication, population health, patient education, digital safety governance and healthcare communication systems rather than direct clinical diagnosis. The chapter title should therefore be understood broadly. AI can support healthcare systems by helping clinicians and public health teams understand the misinformation narratives patients may encounter, prepare evidence-informed responses and anticipate concerns that may influence help-seeking, adherence or vaccination behaviour.

Clinical interpretation remains essential. Social media trends do not necessarily reflect clinical testing data, actual disease burden or population-level risk. Online discussion volume may reflect media attention, political debate, platform algorithms or organised campaigns. Therefore, AI-supported digital surveillance should be interpreted with public health professionals and clinical experts, alongside epidemiological data, service use data and patient experience. National policy frameworks further highlight the importance of integrating AI within broader healthcare governance systems. The NICE Evidence Standards Framework for Digital Health Technologies provides guidance for evaluating digital and AI-enabled tools to ensure they demonstrate clinical effectiveness, safety and value for the health system [36]. The 10 Year Health Plan for England emphasises the growing role of digital technologies and innovation in transforming healthcare delivery and improving population health outcomes [37]. These initiatives demonstrate that AI implementation must be linked to evidence, safety, governance and evaluation.

Despite technological progress, AI cannot replace expert interpretation of medical information. Healthcare professionals remain essential in distinguishing between misinformation, uncertainty, emerging evidence and legitimate scientific debate. Without appropriate oversight, AI models may misclassify evolving medical evidence or patient concerns as misinformation, which could undermine trust. Digital health literacy is also essential, but it should be integrated into the governance framework rather than treated as a separate educational activity. AI-derived public concern signals can identify what people are asking, which groups may be most exposed, which information voids are emerging and which misconceptions require tailored communication [12,14,15].

In practice, social listening outputs can be translated into literacy interventions through a structured pathway: detection of recurring public questions, expert validation of the evidence, segmentation by audience and language, co-production of accessible messages, delivery through trusted clinical and community channels, and evaluation of comprehension, trust and behaviour. For low-risk misunderstandings, this may involve short explanatory materials, frequently asked questions or signposting to authoritative sources. For high-risk misinformation, such as false cancer-treatment claims or unsafe medicine advice, communication should be clinician-reviewed, rapidly disseminated and linked to platform escalation where appropriate. This approach connects AI monitoring directly to targeted digital health literacy, rather than relying on generic public education campaigns [15,38].

## 12. Cost-Benefit and Implementation Considerations

A further limitation in the current evidence base is the limited discussion of cost, resource use and implementation feasibility. AI systems require investment in data access, model development, computing infrastructure, cybersecurity, staff training, clinical review, legal oversight and ongoing evaluation. Larger models and multimodal systems may be expensive and may require technical expertise that some healthcare organisations do not possess. Potential benefits include earlier detection of harmful narratives, improved targeting of public health messages, reduced manual monitoring workload and better understanding of public concerns. However, these benefits must be demonstrated rather than assumed. A technically accurate system may still have limited value if it does not improve trust, reduce exposure to harmful misinformation or support better health behaviours. Decision-makers should consider the severity of the misinformation risk, the expected improvement over existing monitoring systems, the cost of development and maintenance, the workforce required for human review, privacy risks, equity implications and public acceptability. In some cases, a simpler social listening system with expert review may be more appropriate than a complex deep learning system. In other contexts, such as multilingual misinformation during a public health emergency, more advanced AI tools may be justified.

## 13. Future Research Directions

Future research should prioritise methodological rigour, real-world evaluation and governance. First, there is a need for multilingual, cross-platform and topic-diverse datasets. These should include not only COVID-19 but also cancer, chronic disease, reproductive health, nutrition, mental health, vaccination and alternative treatments. Second, studies should include external validation across platforms, languages, populations and time periods. A model trained on one platform should not be assumed to perform effectively elsewhere.

Third, future research should examine annotation quality, class imbalance, error costs and whether datasets distinguish veracity from intent. False positives and false negatives should be assessed in relation to clinical risk, public trust, health equity and scientific debate. Fourth, AI model evaluation should move beyond technical performance metrics. Studies should assess whether AI-supported misinformation governance improves public health communication, tailored digital health literacy, vaccination behaviour, treatment adherence, help-seeking or trust in healthcare systems. Fifth, research should examine human-in-the-loop systems, escalation pathways, iterative feedback loops and audit mechanisms. AI systems should be studied as sociotechnical interventions involving people, institutions, technologies and governance processes. Sixth, future work should assess implementation costs and resource requirements. Public health organisations require evidence on affordability, scalability and value before adopting AI systems at scale.

## 14. Conclusion

AI technologies are increasingly being applied to detect, monitor and counter health misinformation across digital platforms. Evidence suggests that machine learning, natural language processing, deep learning, multimodal systems and social media surveillance can support the identification of misleading health narratives and emerging misinformation trends. However, current evidence does not show that AI models can reliably distinguish intentional disinformation from unintentional misinformation, because most studies classify content by veracity or misleadingness rather than intent. The effectiveness of these technologies depends on addressing dataset bias, platform-specific development, restricted linguistic generalisability, uncertain annotation validity, limited external validation and insufficient evidence of real-world public health benefit. Accuracy and F1-score are useful technical metrics, but they do not demonstrate whether AI improves trust, health literacy, vaccination uptake, treatment adherence or public health behaviour. AI should therefore be viewed as a supportive

tool within an iterative, risk-stratified governance system rather than as a standalone solution. Effective responses require healthcare expertise, public health interpretation, transparent oversight, privacy protection, platform accountability, human review matched to clinical risk, tailored digital health literacy and evaluation of real-world outcomes.

## Key Points

• Artificial intelligence can support health misinformation governance by detecting misleading content, identifying emerging narratives and informing public health communication.

• Current AI approaches remain limited by dataset bias, platform-specific training data, weak external validation and restricted generalisability across languages, populations and health topics.

• AI-supported social media surveillance should be interpreted alongside clinical, epidemiological and public health expertise because online discussion volume does not necessarily reflect disease burden or clinical risk.

• Effective governance requires iterative feedback between detection, validation, interpretation, communication response, platform accountability and evaluation, with human oversight matched to the clinical risk of the content.

• Future research should prioritise multilingual datasets, cross-platform validation, clearer differentiation between misinformation and disinformation, human-in-the-loop review, privacy-preserving methods, cost-benefit analysis and evaluation of public health outcomes.

## Availability of Data and Materials

Not applicable.

## Author Contributions

HHA conceptualised the article, drafted the manuscript, contributed to literature interpretation and critical revision. AAA contributed to the literature interpretation, critical appraisal of the evidence, refinement of the conceptual and governance arguments, proofreading and critical review of the manuscript for intellectual content and editing of the manuscript. Both authors reviewed and approved the final manuscript. Both authors have participated sufficiently in the work and agreed to be accountable for all aspects of the work.

## Ethics Approval and Consent to Participate

Not applicable.


## Acknowledgment

Not applicable.

## Funding

This research received no external funding.


## Conflicts of Interest

The authors declare no conflicts of interest.

## References


[1] Judijanto L, Putra HN, Zani BN, Hasyim DM, Muntasir M. e-Health and digital transformation in increasing accessibility of health services. Journal of World Future Medicine, Health and Nursing. 2024; 2: 119–132.

[2] Pittrow LBRHL. Medical misinformation on social media: assessing the ethical responsibilities of healthcare professionals, social media platforms, and users in combating the spread of false medical information [Doctoral dissertation]. Vilniaus Universitetas. 2025.

[3] World Health Organization. Managing the COVID-19 infodemic: promoting healthy behaviours and mitigating the harm from misinformation and disinformation. Joint statement by WHO, UN, UNICEF, UNDP, UNESCO, UNAIDS, ITU, UN Global Pulse and IFRC. 2020. Available at: https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation (Accessed: 6 March 2026).

[4] Islam MS, Sarkar T, Khan SH, Mostofa Kamal AH, Hasan SMM, Kabir A, et al. COVID-19-Related Infodemic and Its Impact on Public Health: A Global Social Media Analysis. The American Journal of Tropical Medicine and Hygiene. 2020; 103: 1621–1629. https://doi.org/10.4269/ajtmh.20-0812

[5] Roozenbeek J, Schneider CR, Dryhurst S, Kerr J, Freeman ALJ, Recchia G, et al. Susceptibility to misinformation about COVID-19 around the world. Royal Society Open Science. 2020; 7: 201199. https://doi.org/10.1098/rsos.201199

[6] Su Z, McDonnell D, Wen J, Kozak M, Abbas J, Šegalo S, et al. Mental health consequences of COVID-19 media coverage: the need for effective crisis communication practices. Globalization and Health. 2021; 17: 4. https://doi.org/10.1186/s12992-020-00654-4

[7] Eysenbach G. How to Fight an Infodemic: The Four Pillars of Infodemic Management. Journal of Medical Internet Research. 2020; 22: e21820. https://doi.org/10.2196/21820

[8] Cinelli M, Quattrociocchi W, Galeazzi A, Valensise CM, Brugnoli E, Schmidt AL, et al. The COVID-19 social media infodemic. Scientific Reports. 2020; 10: 16598. https://doi.org/10.1038/s41598-020-73510-5

[9] UK Government. National AI Strategy. 2022. Available at: https://www.gov.uk/government/publications/national-ai-strategy (Accessed: 7 March 2026).

[10] NHS England. Planning and implementing real-world artificial intelligence (AI) evaluations: lessons from the AI in Health and Care Award. 2024. Available at: https://www.england.nhs.uk/long-read/planning-and-implementing-real-world-ai-evaluations-lessons-from-the-ai-in-health-and-care-award/ (Accessed: 7 March 2026).

[11] UK Government. Online Safety Act 2023. 2023. Available at: https://www.legislation.gov.uk/ukpga/2023/50/contents (Accessed: 7 March 2026).

[12] Purnat TD, Vacca P, Czerniak C, Ball S, Burzo S, Zecchin T, et al. Infodemic Signal Detection During the COVID-19 Pandemic: Development of a Methodology for Identifying Potential Information Voids in Online Conversations. JMIR Infodemiology. 2021; 1: e30971. https://doi.org/10.2196/30971

[13] Boatman D, Starkey A, Acciavatti L, Jarrett Z, Allen A, Kennedy-Rea S. Using Social Listening for Digital Public Health Surveillance of Human Papillomavirus Vaccine Misinformation Online: Exploratory Study. JMIR Infodemiology. 2024; 4: e54000. https://doi.org/10.2196/54000

[14] White B, Cabalero I, Nguyen T, Briand S, Pastorino A, Purnat TD. An Adaptive Digital Intelligence System to Support Infodemic Management: The WHO EARS Platform. Studies in Health Technology and Informatics. 2023; 302: 891–892. https://doi.org/10.3233/SHTI230296
[15] Liu T, Xiao X. A Framework of AI-Based Approaches to Improving eHealth Literacy and Combating Infodemic. Frontiers in Public Health. 2021; 9: 755808. https://doi.org/10.3389/fpubh.2021.755808
[16] Wang Z, Yin Z, Argyris YA. Detecting Medical Misinformation on Social Media Using Multimodal Deep Learning. IEEE Journal of Biomedical and Health Informatics. 2021; 25: 2193–2203. https://doi.org/10.1109/JBHI.2020.3037027
[17] Wang J, Wang X, Yu A. Tackling misinformation in mobile social networks a BERT-LSTM approach for enhancing digital literacy. Scientific Reports. 2025; 15: 1118. https://doi.org/10.1038/s41598-025-85308-4
[18] Zhang Y, Zong R, Shang L, Yue Z, Zeng H, Liu Y, et al. 'Tripartite Intelligence: Synergizing Deep Neural Network, Large Language Model, and Human Intelligence for Public Health Misinformation Detection (Archival Full Paper)', 2024 ACM Collective Intelligence Conference. Boston, MA, USA, June 27–28, 2024. Association for Computing Machinery: New York, NY, USA. 2024.
[19] Du J, Preston S, Sun H, Shegog R, Cunningham R, Boom J, et al. Using Machine Learning-Based Approaches for the Detection and Classification of Human Papillomavirus Vaccine Misinformation: Infodemiology Study of Reddit Discussions. Journal of Medical Internet Research. 2021; 23: e26478. https://doi.org/10.2196/26478
[20] Zhou Y, Yang Y, Ying Q, Qian Z, Zhang X. 'Multi-modal Fake News Detection on Social Media via Multi-grained Information Fusion', ICMR '23: Proceedings of the 2023 ACM International Conference on Multimedia Retrieval. Thessaloniki, Greece, June 12–15, 2023. Association for Computing Machinery: New York, NY, USA. 2023.
[21] Sharifpoor E, Okhovati M, Ghazizadeh-Ahsaee M, Avaz Beigi M. Classifying and fact-checking health-related information about COVID-19 on Twitter/X using machine learning and deep learning models. BMC Medical Informatics and Decision Making. 2025; 25: 73. https://doi.org/10.1186/s12911-025-02895-y
[22] Valdez D, Soto-Vásquez AD, Montenegro MS. Geospatial vaccine misinformation risk on social media: Online insights from an English/Spanish natural language processing (NLP) analysis of vaccine-related tweets. Social Science & Medicine (1982). 2023; 339: 116365. https://doi.org/10.1016/j.socscimed.2023.116365
[23] Di Sotto S, Viviani M. Health Misinformation Detection in the Social Web: An Overview and a Data Science Approach. International Journal of Environmental Research and Public Health. 2022; 19: 2173. https://doi.org/10.3390/ijerph19042173
[24] Pobiruchin M, Zowalla R, Wiesner M. Temporal and Location Variations, and Link Categories for the Dissemination of COVID-19-Related Information on Twitter During the SARS-CoV-2 Outbreak in Europe: Infoveillance Study. Journal of Medical Internet Research. 2020; 22: e19629. https://doi.org/10.2196/19629
[25] Yan Q, Shan S, Sun M, Zhao F, Yang Y, Li Y. A Social Media Infodemic-Based Prediction Model for the Number of Severe and Critical COVID-19 Patients in the Lockdown Area. International Journal of Environmental Research and Public Health. 2022; 19: 8109. https://doi.org/10.3390/ijerph19138109
[26] Abonizio HQ, Barbon APADC, Rodrigues R, Santos M, Martínez-Vizcaíno V, Mesas AE, et al. How people interact with a chatbot against disinformation and fake news in COVID-19 in Brazil: The CoronaAI case. International Journal of Medical Informatics. 2023; 177: 105134. https://doi.org/10.1016/j.ijmedinf.2023.105134
[27] Cheng Y. Navigating vaccine misinformation: a study on users' motivations, active communicative action and anti-misinformation behavior via chatbots. Online Information Review. 2025; 49: 643–663. https://doi.org/10.1108/OIR-07-2024-0425
[28] Röchert D, Shahi GK, Neubaum G, Ross B, Stieglitz S. The Networked Context of COVID-19 Misinformation: Informational Homogeneity on YouTube at the Beginning of the Pandemic. Online Social Networks and Media. 2021; 26: 100164. https://doi.org/10.1016/j.osnem.2021.100164
[29] Cinelli M, De Francisci Morales G, Galeazzi A, Quattrociocchi W, Starnini M. The echo chamber effect on social media. Proceedings of the National Academy of Sciences of the United States of America. 2021; 118: e2023301118. https://doi.org/10.1073/pnas.2023301118
[30] Pennycook G, McPhetres J, Zhang Y, Lu JG, Rand DG. Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. Psychological Science. 2020; 31: 770–780. https://doi.org/10.1177/0956797620939054
[31] Matias JN. Influencing recommendation algorithms to reduce the spread of unreliable news by encouraging humans to fact-check articles, in a field experiment. Scientific Reports. 2023; 13: 11715. https://doi.org/10.1038/s41598-023-38277-5
[32] Malecki KMC, Keating JA, Safdar N. Crisis Communication and Public Perception of COVID-19 Risk in the Era of Social Media. Clinical Infectious Diseases : an Official Publication of the Infectious Diseases Society of America. 2021; 72: 697–702. https://doi.org/10.1093/cid/ciaa758
[33] Morley J, Cowls J, Taddeo M, Floridi L. Ethical guidelines for COVID-19 tracing apps. Nature. 2020; 582: 29–31. https://doi.org/10.1038/d41586-020-01578-0
[34] Leslie D. Tackling COVID-19 through responsible AI innovation: five steps in the right direction. Harvard Data Science Review. 2020; 2. https://doi.org/10.1162/99608f92.4bb9d7a7
[35] Zenone M, Snyder J, Bélisle-Pipon JC, Caulfield T, van Schalkwyk M, Maani N. Advertising Alternative Cancer Treatments and Approaches on Meta Social Media Platforms: Content Analysis. JMIR Infodemiology. 2023; 3: e43548. https://doi.org/10.2196/43548
[36] National Institute for Health and Care Excellence (NICE). Evidence standards framework for digital health technologies (ECD7). 2022. Available at: https://www.nice.org.uk/corporate/ecd7 (Accessed: 7 March 2026).
[37] NHS England. Fit for the future: 10-year health plan for England. 2025. Available at: https://www.england.nhs.uk/long-term-plan/ (Accessed: 7 March 2026).
[38] Antoliš K. Education against disinformation. Interdisciplinary Description of Complex Systems. 2024; 22: 71–83. https://doi.org/10.7906/indecs.22.1.4