Bixonimania and the epistemic fragility of artificial intelligence in medicine: lessons from a fabricated disease

Lancet Digit Health 2026
Published Online
https://doi.org/10.1016/j.landig.2026.101048

The rapid integration of large language models (LLMs) into clinical and biomedical ecosystems has created unprecedented opportunities for information synthesis, decision support, and patient engagement. Alongside these advances, however, an increasingly concerning phenomenon—the amplification of medical misinformation through generative artificial intelligence (AI) systems—has emerged.[1] The case of bixonimania, a completely fabricated disease that nonetheless circulated in AI-generated outputs as if it were a real clinical entity, serves as a particularly illuminating example.[2]

What makes this phenomenon important is not merely the invention of a fictitious condition but also the speed with which multiple AI systems assimilated it, transforming it into seemingly coherent medical knowledge. Structured symptom profiles, mechanistic explanations, and even diagnostic suggestions were rapidly developed for bixonimania. Importantly, this phenomenon is not simply a case of AI hallucination. Rather, the occurrence can be interpreted as a case of data poisoning and contamination of retrieval systems: two preprints and blog posts on Medium initially introduced the information on bixonimania into the digital ecosystem, after which it was incorporated by chatbots through training data and retrieval-augmented generation systems.[2]

This interpretation is closely linked to empirical research by Liu and colleagues on poisoning attacks in medical LLMs.[3] Adversarial knowledge covertly inserted into retrieval databases, even through nearly imperceptible perturbations, can distort outputs and compromise the reliability of retrieval-augmented generation systems in the medical field. Blatantly incorrect medical information can sometimes be corrected during generation, whereas covert misinformation formulated within the inherent diagnostic ambiguity of medicine is far more likely to evade a model's self-correction mechanisms.[3] The case of bixonimania perfectly illustrates this dynamic: the invented entity was not presented as obviously false but as a plausible new disorder expressed in the formal register of biomedical discourse.

This phenomenon is also linked to the broader issue of hallucinations in foundation models—eg, outputs that are fluent, contextually appropriate, and syntactically convincing without a verified empirical basis.[4] Hallucinations are not occasional errors but, rather, a structural property of probabilistic text-generation systems, especially in domains in which authoritative language strongly influences credibility perceptions. Evaluations of LLMs applied to health care have repeatedly showed their ability to produce clinically coherent narratives containing subtle or overt factual inaccuracies.[1,4] This combination of superficial competence and latent unreliability creates a particularly risky context for clinical implementation.

A further, particularly insidious mechanism is sycophancy—ie, the tendency of LLMs to prioritise agreement with the user's assumptions over factual accuracy. LLMs generally accommodate illogical medical requests, rarely questioning the erroneous assumption contained in the prompt. This behaviour likely stems from reinforcement learning strategies based on human feedback, which reward compliance and conversational alignment.[5] In the context of bixonimania, the implications are clear. If asked, "Could I have bixonimania?", then the question presupposes the existence of the entity; a model prone to sycophancy could therefore reinforce the misunderstanding rather than correct it. The spread of AI-generated misinformation thus represents both a technical problem of model error and a problem of human–machine interactions shaped by behavioural-optimisation dynamics.

Broadly, the bixonimania episode highlights a deeper epistemological tension: the growing difficulty in distinguishing syntactic plausibility from scientific validity in machine-generated medical content. When fictitious entities are expressed through formal biomedical language, they acquire an unintended aura of legitimacy, especially when processed by systems optimised for coherence rather than for truth verification.[4] This vulnerability is not exclusively technological. Modern medicine is indeed founded on layered systems of epistemic trust—ie, peer review, clinical guidelines, institutional validation—but these mechanisms remain vulnerable to dissociation between form and evidential substance.

Such risks are further amplified by peoples' growing reliance on digital tools for interpreting health issues. Cyberchondria is a well documented phenomenon,[6] and emerging evidence suggests that conversational AI could

	Best practice and operational recommendations	Objective
Preprint server screening	Systematically verify whether AI-cited sources have undergone peer review; flag preprints not indexed in MEDLINE or Embase as requiring independent corroboration	Prevent ingestion of unvalidated preprint content (eg, the two preprints that seeded bixonimania) into clinical reasoning workflows
Citation auditing	Cross-reference every AI-generated clinical citation against primary bibliographic databases; treat uncorroborated citations as potentially hallucinated	Detect and intercept fabricated or misattributed references before they enter the scientific record
Training data provenance	Prioritise LLM systems that disclose training data sources and corpus filtering criteria; request vendor documentation of data-curation processes	Enable auditing for potential data-poisoning vulnerabilities, including low-quality web content
AI as decision support only	Use AI outputs as preliminary orientation, not as clinical authority; final decisions remain with qualified clinicians who have reviewed the primary sources	Prevent erroneous clinical interpretations driven by AI-generated content without independent validation
Epistemic transparency with patients	Explicitly inform patients when clinical information has been AI-generated; encourage them to raise AI-sourced diagnoses in direct clinical consultations	Reduce health anxiety and prevent patients from acting on unverified AI-generated diagnostic hypotheses
Cyberchondria monitoring	Identify patients presenting with symptom-search anxiety amplified by AI-generated clinical narratives; provide structured epistemic reassurance	Limit diagnostic anxiety escalation driven by AI systems producing confident-sounding but unverified content
Academic-integrity workflows	Cross-check cited entities against controlled vocabularies (eg, ICD-11, MeSH, SNOMED CT); flag any clinical terms absent from established disease ontologies	Detect and mitigate AI-generated or AI-amplified misinformation in scholarly literature, as illustrated by the *Cureus* retraction[8]
Corpus governance and retrieval integrity for RAG systems	Restrict retrieval corpora in RAG-based clinical tools to indexed, peer-reviewed sources; implement editorial-status filters (eg, flags for retracted papers, expressions of concern, and unverified preprints) before LLM ingestion	Reduce risk of training-data or retrieval-level poisoning by unvetted online content
Agentic AI oversight	Subject agentic clinical AI systems with open-web access to enhanced pre-deployment audits; mandate human-in-the-loop review	Prevent agentic architectures from serving as amplification vectors for fabricated clinical entities
Continuous output monitoring	Continuously audit deployed clinical AI systems for emergence of unverifiable clinical entities; establish retraction protocols analogous to academic publishing	Prevent long-term propagation of AI-generated medical misinformation

AI=artificial intelligence. LLM=large language model. MeSH=Medical Subject Headings. SNOMED CT=Systematized Nomenclature of Medicine–Clinical Terms. RAG=retrieval-augmented generation.

Table: **Case-specific mitigation strategies based on the bixonimania example**

intensify it by producing personalised and contextualised responses that mimic clinical reasoning without any basis in validated evidence.[7] Unlike traditional search engines, LLMs do not merely retrieve information; they synthesise, reinterpret, and personalise it, often constructing narratives that reinforce pre-existing anxieties or generate new diagnostic concerns. Sycophantic reinforcement further exacerbates this process: repeated interactions can lead to increasingly confident confirmations of unfounded diagnostic beliefs, creating a self-reinforcing cycle of health anxiety.

A particularly emblematic example of contamination of the scientific ecosystem occurred on March 30, 2026, when *Cureus* formally retracted a manuscript that cited bixonimania as a real clinical entity.[8] This example shows how AI-generated misinformation can bypass traditional peer-review barriers when embedded in formally correct biomedical language. The incident also highlights the urgency of developing academic-integrity workflows capable of identifying terms absent from controlled vocabularies such as those in ICD-11, MeSH, and SNOMED CT.

Debates regarding agentic AI and retrieval-augmented generation architectures underscore the importance of architectural choices in mitigating or amplifying these risks. A systematic review of agentic AI in clinical settings showed that multiagent systems based on distinct roles, in which specialised agents handle retrieval, synthesis, analysis, and quality control through structured workflows, can significantly reduce hallucination rates. Some investigators observed hallucination reductions to 0–10% in specific clinical-guidance tasks when using strictly governed retrieval-augmented generation pipelines.[9]

These results are particularly instructive when applied to the case of bixonimania. A retrieval-augmented generation system restricted to indexed and peer-reviewed sources would likely have excluded the original preprints and the Medium article, preventing the fictional entity from entering the model's outputs. Conversely, an agent-based system with internet access and autonomous retrieval capabilities could have further amplified the misinformation by repeatedly retrieving and recirculating the same low-quality sources across multiple reasoning stages. Similarly, a weakly governed retrieval-augmented generation corpus, including preprints or unfiltered internet content, would have remained highly vulnerable. These issues are not hypothetical: debates on the safety of clinical AI from the 2025–26 period have increasingly focused on issues of architectural governance.[9]

To move from abstract concerns to concrete operational measures, the table presents mitigation strategies explicitly inspired by the case of bixonimania, focusing on preprint

screening, citation auditing, and training-data traceability, rather than on generic AI safety recommendations.

The practical implications for clinical medicine, including a six-step protocol for responding to individuals presenting with AI-generated diagnostic hypotheses, are described in the appendix (pp 1–2). The appendix (pp 3–4) also illustrates the six-stage propagation pathway of bixonimania, from its initial online fabrication to its subsequent formal retraction. The sequence—fabricated online content → digital dissemination → ingestion into AI models → clinically plausible chatbot outputs → academic citation → retraction—shows how a single fictitious entity can propagate from low-authority internet sources into formal scientific literature via generative systems. Importantly, each stage also represents a potential point of intervention, ranging from corpus governance and preprint screening during the ingestion phase to academic-integrity checks and continuous monitoring after deployment.

Despite these key issues, adopting a technophobic interpretation of the phenomenon would be inappropriate. LLMs and AI systems offer substantial benefits in medicine, including evidence synthesis, educational support, and clinical empowerment. The central question is therefore not whether such systems should be used but rather how their outputs should be interpreted, validated, and governed.

Ultimately, the case of bixonimania serves as a conceptual stress test for contemporary digital health infrastructures. This example shows how rapidly fabricated medical knowledge can emerge, spread, and acquire apparent legitimacy when mediated by generative systems without solid epistemic grounding. Crucially, the problem goes beyond mere hallucination: it involves data poisoning, retrieval from unvalidated sources, sycophantic reinforcement of users' assumptions, and, perhaps most worryingly, the ability of such fabrications to even enter peer-reviewed literature.

The future of digital health will depend on both sophisticated AI and robust epistemic safeguards to govern its use. These safeguards should extend beyond mere technical-mitigation strategies and include regulatory oversight, patient digital literacy, and profound reforms in the training, validation, and auditing processes of AI systems intended for clinical use.

See **Online** for appendix

We declare no competing interests.

Copyright © 2026 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Francesco Branda, Massimo Ciccozzi
f.branda@unicampus.it

Unit of Medical Statistics and Molecular Epidemiology, Università Campus Bio-Medico di Roma, Rome 00128, Italy (FB, MC)

1 Omar M, Sorin V, Wieler LH, et al. Mapping the susceptibility of large language models to medical misinformation across clinical notes and social media: a cross-sectional benchmarking analysis. *Lancet Digit Health* 2026; **8:** 100949.
2 Stokel-Walker C. Scientists invented a fake disease. AI told people it was real. *Nature* 2026; **652:** 559–61.
3 Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. *J Am Med Inform Assoc* 2025; **32:** 605–15.
4 Roustan D, Bastardot F. The clinicians' guide to large language models: a general perspective with a focus on hallucinations. *Interact J Med Res* 2025; **14:** e59823.
5 Rosen KL, Sui M, Heydari K, Enichen EJ, Kvedar JC. The perils of politeness: how large language models may amplify medical misinformation. *NPJ Digit Med* 2025; **8:** 644.
6 Miezah D, Adzakpa R, Bani EMA, Obeng P. Prevalence and associated factors of cyberchondria: a scoping review. *ScientificWorldJournal* 2026; **2026:** 9950027.
7 Vasiliu O, Ciobanu AM. Diagnosis and therapeutic challenges in the exploration of cyberchondria. *Eur Psychiatry* 2025; **68** (suppl 1)**:** S498. (abstr).
8 Banchhor S, Gupta S, Mahendra A. Retraction: Clinical and dermoscopic evaluation of periorbital melanosis and its psychological impact and effect on quality of life: a descriptive study. *Cureus* 2026; **18:** r223.
9 Salehi S, Singh Y, Horst KK, Hathaway QA, Erickson BJ. Agentic AI and large language models in radiology: opportunities and hallucination challenges. *Bioengineering (Basel)* 2025; **12:** 1303.