doi:10.3233/SHTI250219

The Unexpected Harms of Artificial Intelligence in Healthcare: Reflections on Four Real-World Cases

Kerstin DENECKE[a,1] Guillermo LOPEZ-CAMPOS[b], Octavio RIVERA-ROMERO[c,] and Elia GABARRON[d]
[a] *Bern University of Applied Sciences, Bern, Switzerland*
[b] *Wellcome-Wolfson Institute for Experimental Medicine, Queen's University Belfast*
[c] *Department of Electronic Technology, Universidad de Sevilla, Seville, Spain*
[d] *Department of Education, ICT and Learning, Østfold University College, Norway*
ORCiD ID: Kerstin Denecke https://orcid.org/0000-0001-6691-396X, Guillermo Lopez-Campos https://orcid.org/0000-0003-3011-0940, Octavio Rivera-Romero https://orcid.org/0000-0001-7212-9805, Elia Gabarron https://orcid.org/0000-0002-7188-550X

Abstract. *Introduction:* Rapid advances in Artificial Intelligence (AI), especially with large language models, present both opportunities and challenges in healthcare. This article analyzes real-world AI-related harms in healthcare. *Methods:* We selected four recent AI-related incidents from the AIAAIC Repository. *Results:* The incidents discussed include: Whisper's harmful hallucinations; UNOS's algorithm delaying transplants for black patients; the WHO's S.A.R.A.H. chatbot providing inaccurate health information; and Character AI's chatbot promoting disordered eating among teens. *Discussion and conclusion*: These incidents highlight diverse risks, from misinformation to safety concerns, involving both industry and institutional providers. The article emphasizes the need for systematic reporting of AI-related harms, concerns about security, privacy, and ethics, and calls for a centralized health-specific database to enhance patient safety and understanding.

Keywords. Artificial intelligence, Digital technology, Digital health interventions, Adverse events, Patient safety

1. Introduction

Rapid advances in artificial intelligence (AI), particularly with large language models (LLMs) and generative AI, have created both new opportunities and challenges in healthcare. These technologies have demonstrated remarkable capabilities in understanding and generating human language, and have proven highly effective in natural language processing (NLP) tasks like translation [1], summarization [2], classification [3], named entity recognition, and medical question answering [4]. Despite these recent advances, AI has been used in medical informatics for over half a century and has been extensively used for decades in different areas such as the development of

[1] Corresponding Author: Kerstin Denecke. Bern University of Applied Sciences, Quellgasse 21, 2502 Biel/Bienne, Switzerland. Mail: kerstin.denecke@bfh.ch.

clinical decision support systems [5,6]. Along its way, the development of such solutions has not been exempted from challenges and concerns about the development of such approaches. A paradigmatic recent example of these risks were the racial biases in measurements coming from pulse-oximeters and other medical devices [7,8]. While AI integration can increase efficiency and optimization of healthcare processes, concerns remain regarding accuracy, regulatory compliance, privacy and security, human factors, and ethical considerations [9].

This research article aims to document and analyze examples of real-world cases of incidents and harms in healthcare linked to the use of AI, emphasizing the critical importance of recording these events in the scientific literature to better understand their scope and implications, and to guide the development of strategies for mitigating risks and promoting the safe adoption of AI in clinical settings.

2. Methods

We randomly selected four recent AI-related incidents from the AIAAIC Repository: two affecting healthcare professionals and two impacting general population and children. AIAAIC is one of the independent initiatives that focuses on advocating for transparency and openness in AI algorithms. This initiative maintains a repository where harms of AI and algorithmic systems across all sectors are recorded [10]. As of early 2025, it has recorded 1,904 incidents since 2008, with over 100 linked to healthcare. Table 1 summarizes the covered diverse technologies, risks, tasks, and users involved in the four selected cases.

Table 1. Overview of the described cases

Aspects	Case 1	Case 2	Case 3	Case 4
Technology	Speech-to-Text Technology	Algorithm for prioritization	Chatbot	Chatbot
Risk / Safety concern	Hallucinations, Risk for patient safety, Risk for data integrity in health records	Delayed transplants, patient safety	Outdated information, inaccurate information	"Coach" anorexia-like behaviors
Task	Clinical documentation	Decision-making	General health advice	Chat
User	Health professionals	Health professionals	General population	Teenagers

3. Results

3.1. *Case 1: AI transcriptions as a risk for patient safety and data integrity* [11,12]

Whisper, an automatic speech recognition system trained on 680,000 hours of multilingual data, has been found to generate false text, sometimes producing entire sentences that were not present in the original audio [13]. These "hallucinations" can involve harmful content, such as racist comments, violent rhetoric, and fabricated medical treatments, like a non-existent drug called "hyperactivated antibiotics." A study involving 13,140 audio segments found that 1.4% contained hallucinations, with nearly 40% being harmful or concerning [14]. While no direct patient harm has been reported, inaccurate clinical transcripts pose risks to patient safety. Although Whisper transcribed spoken content correctly, it added false information, including violence, inaccurate associations, and false authority [14]. Despite OpenAI's warnings against using Whisper in high-risk areas, it is being adopted in healthcare, raising concerns about patient safety, medical record integrity, and confidentiality.

3.2. *Case 2: Algorithm delays transplants for black patients and youth* [15,16]

Several incidents highlight biases in algorithms used to prioritize organ transplant patients. In April 2023, the UNOS's UNet algorithm in the U.S. was found to unfairly delay kidney transplants for Black patients by overestimating their kidney function, leading to longer wait times [15]. Similarly, the Transplant Benefit Score algorithm in the UK, introduced in 2018, assigned lower scores to younger patients, reducing their chances of receiving a liver transplant [16]. These incidents highlight the importance of analyzing biases in AI algorithms used in healthcare. Healthcare professionals must be aware of these biases to prevent discriminatory outcomes. In one case, this bias resulted in a patient waiting over five years for a kidney transplant.

3.3. *Case 3. WHO chatbot provides inaccurate health information* [17]

In April 2024 the World Health Organization (WHO) released S.A.R.A.H. [18], a digital health promoter based on ChatGPT3.5, designed to provide guidance on topics such as mental health, healthy eating or quitting smoking amidst a growing shortage of healthcare workers. However, a media report shortly after the official release of the tool that the system failed to provide updated and accurate information. The WHO acknowledged these limitations, noting that S.A.R.A.H. is still a work in progress and often directs users to its website or healthcare providers. The incident highlights concerns about the accuracy and timeliness of AI in healthcare. S.A.R.A.H. includes a disclaimer that its responses do not reflect WHO's views and are not guaranteed to be accurate. Similar issues previously led to the shutdown of an eating disorder support chatbot [19]. As of this writing, no further studies or updates on S.A.R.A.H. have been found, and it remains unclear whether its performance has improved.

3.4. *Case 4: Character AI encourages kids to engage in disordered eating* [20]

Character AI, a platform hosting chatbot personas [21], faced media exposure after some of its chatbots, like "4n4 Coach" (a twist on "ana", the online nickname for anorexia), promoted disordered eating behaviors among teens. These bots encouraged dangerously low-calorie diets (e.g., 900–1,200 calories daily), meal skipping, and excessive exercise, engaging nearly 14,000 users [20] and highlighting lapses in content moderation, age restrictions, and ethical oversight. While no direct harm has been proven, exposure to pro-anorexia content can negatively impact adolescents' body image and eating behaviors [22,23]. It is worth mentioning that at the time of writing this article, the "4n4 Coach" no longer appears in search results. However other pro-anorexia bots remain active, some with over 1,000 users. This incident highlights the risks of unregulated AI, especially for vulnerable youth, raising concerns about eating disorders and mental health. As of early 2025, to our knowledge, there are no publications indexed in a leading health literature database (e.g., PubMed) referencing this case.

4. Discussion

The described AI technologies target different audiences, from the general public (WHO chatbot), teens (Character AI), to healthcare professionals (transplant algorithms, speech-to-text tools). The risks also vary, with one case directly affecting patient safety and others posing potential harm. These cases involve both industry and institutional providers, but only one had scientific literature support [14].

The growing use of AI in health drives both research and industry. Research prioritizes clinical effectiveness and best practices [24], while industry operates in both regulated and unregulated spaces, where risks may go unreported. Furthermore, no centralized database exists for AI-related healthcare incidents, and existing repositories, such as AIAAIC [10], AI Incident Database [25] and the OECD AI Incident Monitor [26] rely on voluntary reporting. In the USA, AI-related medical device issues may be found in the FDA's MAUDE database [27]. However, the absence of a dedicated health-specific database limits risk understanding, affecting patient safety and public trust.

Experts highlight concerns regarding misinformation, security, privacy, ethics, and liability [9,28-31]. The presented real-world cases align with these concerns: Whisper's hallucinations (case 1) involve misinformation, biased transplant algorithms (case 2) reinforce discrimination, and cases 3 and 4 demonstrate AI-related safety risks. Despite warnings from AI developers, healthcare institutions continue adopting AI to address staff shortages and streamline processes, often overlooking risks. Without regulations, this trend is likely to worsen. To mitigate harm, WHO has issued AI ethics and governance guidance [32], emphasizing participatory design, risk prediction, and regulatory enforcement. Scientific documentation of real-world AI incidents is crucial for transparency and responsible AI integration [33].

This study has several limitations. The cases were randomly selected rather than using a systematic methodology, limiting the generalizability and representativeness of the findings. While our goal was to highlight AI-related harms in healthcare and encourage systematic reporting, this approach may not capture the full scope or frequency of such incidents. Additionally, reliance on anecdotal examples prevents quantifying the prevalence or severity of these harms, emphasizing the need for future research with more rigorous methods. Future research could systematically analyze documented incidents of harmful AI applications in healthcare available in repositories.

5. Conclusion

The real-world cases presented in this paper highlight the significant risks and ethical challenges associated with the use of AI in healthcare, including transcription hallucinations, biased transplant algorithms, inaccurate health information, and the promotion of disordered eating. These incidents, often underreported in scientific literature, underscore the urgent need for monitoring and systematic reporting of AI-related harms, while emphasizing the importance of transferring knowledge from non-scientific media to the scientific community to address these challenges effectively.

References

[1] Wang W, Yang Z, Gao Y, et al. Transformer-Based DirectHidden Markov Model for Machine Translation. Proceedingsof the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, Online: Association for Computational Linguistics. 2021:23-32.

[2] Moro G, Ragazzi L, Valgimigli L, et al. Efficient Memory-Enhanced Transformer for Long-Document Summarization in Low-Resource Regimes. Sensors. 2023;23:3542.

[3] Dai X, Chalkidis I, Darkner S, et al. Revisiting Transformer-based Models for Long Document Classification. In Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi, United Arab Emirates Association for Computational Linguistics. 2022:7212-7230.

[4] Yang X, PourNejatian N, Shin HC, et al. Gatortron: A large language model for clinical natural language processing. MedRxiv. 2022:2022-02.

[5] Shortliffe EH, Axline SG, Buchanan BG, et al. An artificial intelligence program to advise physicians regarding antimicrobial therapy. Comput Biomed Res. 1973;6(6):544-60.

[6] Hand DJ. Artificial intelligence and medicine: discussion paper. J R Soc Med. 1987;80(9):563-5.

[7] Hidalgo DC, Olusanya O, Harlan E. Critical care trainees call for pulse oximetry reform. Lancet Respir Med. 2021;9(4):e37.

[8] Valbuena VSM, Merchant RM, Hough CL. Racial and Ethnic Bias in Pulse Oximetry and Clinical Outcomes. JAMA Intern Med. 2022;182(7):699-700.

[9] Denecke K, May R, Rivera-Romero O. Transformer Models in Healthcare: A Survey and Thematic Analysis of Potentials, Shortcomings and Risks. J Med Syst. 2024;48(1):23.

[10] AIAAIC (AI AaAIaC. AIAAIC 2024. Available from: https://www.aiaaic.org/

[11] AIAAIC. Study: Whisper AI transcription service invents medical treatments 2024. Available from: https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/study-whisper-ai-transcription-invents-medical-treatments

[12] AIAAIC. Study: Whisper AI speech recognition creates violent hallucinations 2024. Available from: https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/study-whisper-ai-speech-recognition-creates-violent-hallucinations

[13] Burke G, Schellmann H. Researchers say an AI-powered transcription tool used in hospitals invents things no one ever said 2024. Available from: https://apnews.com/article/ai-artificial-intelligence-health-business-90020cdf5fa16c79ca2e5b6c4c9bbb14

[14] Koenecke A, Seo A, Choi G, et al. Careless Whisper: Speech-to-Text Hallucination Harms. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24) Association for Computing Machinery, New York, NY, USA. 2024:1672-1681.

[15] AIAAIC. Black man sues UNOS over kidney transplant algorithm racial bias 2023. Available from: https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/black-man-sues-unos-over-kidney-transplant-algorithm-racial-bias

[16] AIAAIC. Algorithm delays young peoples' liver transplants 2023. Available from: https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/algorithm-delays-young-peoples-liver-transplants

[17] AIAAIC. WHO chatbot provides inaccurate health information 2024. Available from: https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/who-chatbot-provides-inaccurate-health-information

[18] World Health Organization. S.A.R.A.H, a Smart AI Resource Assistant for Health 2024. Available from: https://www.who.int/campaigns/s-a-r-a-h

[19] Wells K. An eating disorders chatbot offered dieting advice, raising fears about AI in health 2023. Available from: https://www.npr.org/sections/health-shots/2023/06/08/1180838096/an-eating-disorders-chatbot-offered-dieting-advice-raising-fears-about-ai-in-hea

[20] AIAAIC. Character AI encourages kids to engage in disordered eating 2024. Available from: https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/character-ai-encourages-kids-to-engage-in-disordered-eating

[21] Character Technologies I. https://character.ai/ 2024. Available from: https://character.ai/

[22] Mento C, Silvestri MC, Muscatello MRA, et al. Psychological Impact of Pro-Anorexia and Pro-Eating Disorder Websites on Adolescent Females: A Systematic Review. Int J Environ Res Public Health. 2021;18(4).

[23] Rodgers RF, Lowy AS, Halperin DM, et al. A Meta-Analysis Examining the Influence of Pro-Eating Disorder Websites on Body Image and Eating Pathology. Eur Eat Disord Rev. 2016;24(1):3-8.

[24] Embi PJ. Algorithmovigilance-Advancing Methods to Analyze and Monitor Artificial Intelligence-Driven Health Care for Effectiveness and Equity. JAMA Netw Open. 2021;4(4):e214622.

[25] AI Incident Database. New Incident Report 2024. Available from: https://incidentdatabase.ai/apps/submit/

[26] OECD.AI Policy Observatory. G7 reporting framework – Hiroshima AI Process (HAIP) international code of conduct for organizations developing advanced AI systems 2025. Available from: https://transparency.oecd.ai/

[27] FDA U.S. Food & Drug Administartion. Manufacturer and User Facility Device Experience (MAUDE) Database 2025. Available from: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/search.cfm

[28] Denecke K, May R, Rivera Romero O. Potential of Large Language Models in Health Care: Delphi Study. J Med Internet Res. 2024;26:e52399.

[29] Guardado S, Mynolopoulou V, Bansi J, et al. An exploratory study on the opportunities and usefulness of patient-generated health data as a tool in the care of people with Multiple Sclerosis. Methods of Information in Medicine. 2023.

[30] Rani S, Kumari A, Ekka SC, et al. Perception of Medical Students and Faculty Regarding the Use of Artificial Intelligence (AI) in Medical Education: A Cross-Sectional Study. Cureus. 2025;17(1):e77514.

[31] Stroud AM, Curtis SH, Weir IB, et al. Physician Perspectives on the Potential Benefits and Risks of Applying Artificial Intelligence in Psychiatric Medicine: Qualitative Study. JMIR Ment Health. 2025;12:e64414.

[32] World Health Organization. Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models 2024. Available from: https://www.who.int/publications/i/item/9789240084759

[33] Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Med Res Methodol. 2022;22(1):287.