MEDINFO 2025 — Healthcare Smart × Medicine Deep
M.S. Househ et al. (Eds.)

doi:10.3233/SHTI250992

The Unintended Harm of Artificial Intelligence (AI): Exploring Critical Incidents of AI in Healthcare

Kerstin DENECKE[a,1] Guillermo LOPEZ-CAMPOS[b] and Richard MAY[c]
[a] *Bern University of Applied Sciences, Bern, Switzerland*
[b] *Wellcome-Wolfson Institute for Experimental Medicine, Queens University Belfast, Belfast, UK*
[c] *Harz University of Applied Sciences, Wernigerode, Germany*
ORCiD ID: Kerstin Denecke https://orcid.org/0000-0001-6691-396X, Guillermo Lopez-Campos https://orcid.org/0000-0003-3011-0940, Richard May https://orcid.org/0000-0001-7186-404X

Abstract. Artificial intelligence (AI) has been utilized in healthcare for years, presenting various risks. However, there is a gap in understanding AI incidents, particularly their impacts and associated risks. This study provides an overview of AI-related incidents in healthcare, focusing on the technologies involved, their contexts (e.g., medical conditions or treatments), and the effects of these incidents. We explored cases from the AI Algorithmic and Automation Incidents and Controversies (AIAAIC) repository, identifying 82 health-related incidents, including 14 cases with high and 1 case with critical severity. Our findings highlight significant concerns about AI misbehavior affecting patient safety, especially in medical monitoring, prediction, and diagnosis systems, including those not specifically designed for these functions (e.g., ChatGPT). The severity of these incidents underscores the need to enhance the robustness and reliability of AI systems, impacting trust in such technologies.

Keywords. Artificial Intelligence, Incidents, Risks, Harm, Patient safety, Healthcare

1. Introduction

These days, the application of Artificial Intelligence (AI) in healthcare spans across multiple areas ranging from image segmentation and classification, clinical decision support systems, drug design and to the more recent development of conversational agents (i.e., chatbots) [1,2]. Despite AI has been used in health informatics for decades, as it is exemplified by its inclusion as a MeSH term in 1986 (https://meshb.nlm.nih.gov/record/ui?ui=D001185), recent developments and applications in many other areas have re-ignited the interest and development of AI applications for health. We have witnessed an explosion in the use of AI tools that affect multiple aspects in our societies. However, the use of AI applications includes, not

[1] Corresponding Author: Kerstin Denecke, Institute Patient-centered Digital Health, Bern University of Applied Sciences, Quellgasse 21, 2502 Biel, Switzerland, kerstin.denecke@bfh.ch.

surprisingly, a variety of risks, in particular regarding safety – leading to several novel concepts such as digitalovigilance and algorithmovigilance to investigate the effects of AI and other digital technologies in healthcare [3–5]. Therefore, in parallel with their implementation, there has been an increasing interest in monitoring and ensuring their proper use and behavior. Consequently, several databases and tools have been developed to track incidents where AI technologies are involved. Examples of these resources are the *Artificial Intelligence Incident Database* (*AIID*, https://incidentdatabase.ai/) or the *AI, Algorithmic and Automation Incidents and Controversies repository* (*AIAAIC*, https://www.aiaaic.org). Both examples offer information about incidents involving a broad range of affected domains (e.g., arts, automotive, healthcare), stakeholders or severity [6,7].

We argue that there is still a gap in the current understanding of AI incidents, in particular concerning the impacts of these incidents and associated risks in a safety-critical environment like healthcare. To address this issue, we aim to provide an overview of AI-related incidents in healthcare, focusing on the technologies involved, the context where they are involved (e.g., medical conditions or treatments) and the effects associated with these incidents.

2. Method

To address our goal, we did a case survey based on incidents related to AI in healthcare mined from the *AIAAIC* repository, relying on established guidelines for mining datasets in software engineering [8]. We combined two methods: (1) manually analyzing data and (2) automatically modeling topics building on Latent Dirichlet Allocation (LDA) [9].

Dataset. The AIAAIC repository comprises a catalog of media reports on real-world AI incidents, including newspaper articles or technical reports of organizations. With more than 1’800 entries it is currently one of the largest open-access datasets related to such incidents [6,7], offering detailed classifications of them (i.e., 16 categories), such as domain, systems, or risks. Interestingly, it comprises a large part of incidents that are originally listed in other databases, such as *Where in the World is AI*, the *AI Vulnerability Database*, or the *AI Incident Database* [6,11], leading to a smaller selection bias. Typically, a URL is given to every entry, providing the option to access corresponding short descriptions of the incidents and associated media reports. Generally, the AIAAIC repository is known for its high data quality [7], which is emphasized by scientific work, already building from different perspectives on its data and classifications, e.g., related to risks such as ethics [10] or safety and security [11], or domains such as finance or critical infrastructure [12].

Data selection. We downloaded the Excel spreadsheet on the AIAAIC website (https://www.aiaaic.org/aiaaic-repository, accessed October 02, 2024). To select data, we only considered incidents of the last decade (2015–2024) and reports that are still available on the Internet and describe an actual incident related to the health domain (i.e., excluding issues that might only lead to incidents). Applying these criteria, one author (RM) initially reduced the dataset to 1’053 domain-independent incidents and, after filtering the domain, to 97 incidents related to healthcare. After screening, a total of 82 cases were considered for analysis (all data will be made available after acceptance).

Data analysis. For the manual data analysis, we focused on five repository categories (i.e., ID, title, year, working area, issues), additionally considering two more categories with partly missing data for supporting further findings (i.e., harms).

Furthermore, we created two categories on our own which are based on the short descriptions and full texts of the incidents: origin (i.e., misbehavior vs. misuse) and severity. The latter comprises four potential classifications, including low (no immediate or serious impact on health), moderate (potentially negative impact on health), high (direct risk to the health or life), and critical (serious and irreversible health issues).

To clearly identify the AI systems' underlying technology, RM applied an automated topic modeling algorithm (i.e., LDA). We used the short descriptions as the basis for the LDA algorithm (i.e., a Python script with *NLTK*, *stop_words*, *gensim*). After text pre-processing (i.e., removing hyperlinks, tokenization, lemmatization, vectorization), LDA was applied by experimenting with parameters to find a suitable coherence value. Specifically, the number of topics k and iterations i was varied (k between eight and 25; i between 100 and 500 with steps of 100) until we reached a coherence value around 0.6 ($k = 14$, $i = 200$), indicating reliable results [13]. The choice of k and i was based on coherence trends: lower k led to overly broad topics, while higher k caused fragmentation. Similarly, increasing i beyond 200 showed only marginal coherence improvements. Via open-card sorting, the LDA topics were labeled with suitable classifications; a final validation of all incidents by the author RM led to a final number of 13 technology topics. Note, that we are aware of potential threats to the internal validity posed by the single-author analysis. However, we argue that the author has extensive experience with similar datasets [11] and only clear trends that were discussed by all authors were finally considered.

3. Results

The 82 incidents originated from reports between 2015 and 2024, with an average of 8.2 reported incidents per year (minimum: 1 in 2017 and maximum 23 in 2021). Peaks are in 2020 and 2021 with 17 and 23 reports per year. 96% of the reports (79/82) reflect a misbehavior of an AI-based tool while only 4% (3/82) refer to misuse of an AI-based tool by a human. 12 cases were dealing with COVID-19, two related to liver transplant, 3 about kidney diseases, cancer, autism, depression and others. The three cases of misuse relate to 1) a student who uses ChatGPT to dope a health insurance (AIAAIC0512), 2) an AI-generated deepfake where a TV moderator advertises a libido booster (AIAAIC0213) and 3) Google wrongly flagged medical images a parent took of his son's groin as child sexual abuse material (AIAAIC1095). Figure 1 shows the involved technologies. Predictive systems account for the largest share of reported incidents at 18%, closely followed by conversational agents at 17%. Monitoring systems account for 12% of reports and diagnostic systems for 11%.

The severity of the incidents was low (36/82, 44%), moderate (31/82, 38%) and high (14/82, 17%). One incident was classified as critical, i.e., irreversible health issues. This incident concerned robotic surgery that was linked to 144 deaths. Incidents of high severity concerned prediction systems (4/14), a conversational agent (1/14), a monitoring system (1/14), diagnosis systems (3/14), recommender systems (2/14), resource allocation systems (2/14) and robotics (1/14).

Most AI applications in our dataset are related to information gathering or provision (29/82, 35%) and disease assessment (14/82, 17%). Other application areas are disease prediction, disease monitoring, health monitoring, and organ function assessment with 5-6% of the cases each. Applications that occur either once or twice include: documentation, ingredient analysis, fertility assessment, organ resource allocation, care resource allocation, drug abuse, surgical operation, person identification, organ transplantation, drug recommendation, drug distribution, health prediction, and healthcare needs prediction.

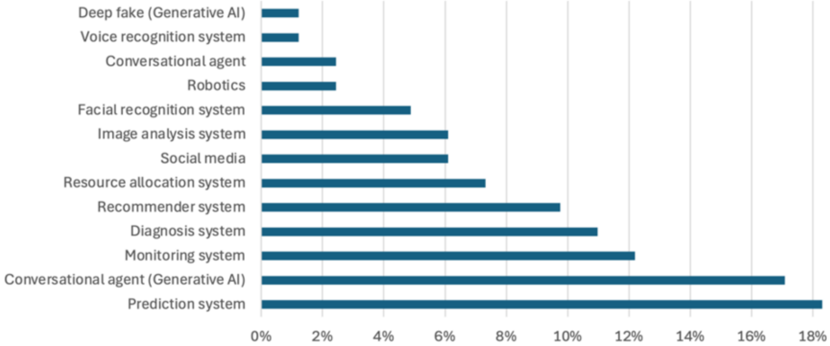

Figure 1. Technologies involved in the reported incidents.

Table 1 shows the topics of the reported incidents. Most incidents are related to accuracy and safety. Misinformation, ethics, security and privacy are other important topics of the incidents. An example of an incident with a conversational agent is an incident on ChatGPT that provided inaccurate medication query responses (AIAAIC1262). There were also several incidents that caused privacy loss as external harm. Others created harm in individuals, such as leading to loss of life, health deterioration, damage in physical health and safety. Internal harms concern litigation, system or algorithm updates, or regulatory inquiries, reputational damage.

Table 1. Topics of the reported incidents.

Issue / risk	Amount (%)
Accuracy/reliability	67 (83%)
Safety	64 (78%)
Mis/disinformation	31 (38%)
Ethics/values	26 (32%)
Security	25 (30%)
Privacy	24 (29%)
Fairness	21 (26%)
Bias/discrimination	18 (22%)
Dual/multi-use	6 (7%)

Issue / risk	Amount (%)
Scope creep/normalisation	4 (5%)
Appropriateness/need	3 (4%)
Robustness	2 (2%)
Confidentiality	2 (2%)
Environment	1 (1%)
Governance	1 (1%)
Marketing	1 (1%)
Anthropomorphism	1 (1%)
Accountability	1 (1%)

4. Discussion

Our data raises serious concerns about the misbehavior of AI in relation to patient safety when AI systems are used in healthcare, particularly related to medical monitoring, prediction, and diagnosis systems – including those that are not necessarily designed for it, but still have such functions integrated (e.g., ChatGPT). A particularly relevant aspect identified in our analyses was the severity of the incidents. Although most of them were catalogued as low or moderate, we were able to identify 18% of them had a high or critical severity, i.e., they had an impact on patient care and even led to loss of lives. For example, liver transplants in young people were delayed by an AI algorithm; a kidney disease care algorithm failed due to racial bias, or an AI risk prediction algorithm identified wrongly high-risk patients. Interestingly, in terms of the medical conditions found in our analyses, COVID-19 was identified as the main condition associated with misbehavior of AI tools. This is not surprising given that incidents peaked in 2020 and 2021 during the COVID-19 pandemic, implicating that patients tried to inform or treat themselves via AI tools. Note, although not every incident is necessarily critical, even a few incidents are unacceptable due to their major impact on patient safety. So, building on our results, the prevalence of AI misbehavior incidents underscores a strong need for improving the robustness and reliability of AI systems, including implications on trust in such systems. For AI to be effectively integrated into healthcare, it is essential to build and maintain trust among healthcare providers and patients and restrict and/or strictly validate misleading functionalities of generative AI. Incidents involving privacy loss highlight the importance of strong data protection measures. We note that privacy issues may also be related to AI security measures (e.g., to protect models from adversarial attacks). Healthcare data is particularly sensitive, and breaches can have critical consequences for patient confidentiality and trust.

In recent years related work, namely by Burema et al. [10] and May et al. [11], has been published; but they either lacked the health-oriented focus or the depth of our analyses (i.e., in number of reports analyzed or domain-specific aspects). A common area studied with those studies were the topics of the incident where our results corroborated their detection of safety and accuracy/reliability aspects identified in our analyses. In this regard it is important to note that reported incidents were usually labelled with more than one topic and in particular both accuracy/reliability and safety were present in a majority of the reported incidents, e.g., a prediction system was related to the Epic systems sepsis prediction model with risks regarding both safety and accuracy (AIAAIC0657). We conclude that there is an immense need to foster research on the risks of AI-based solutions for patient safety.

References

[1] Kasula BY. Advancements in AI-driven Healthcare: A Comprehensive Review of Diagnostics, Treatment, and Patient Care Integration. International Journal of Machine Learning for Sustainable Development. 2024; 6:1–5.

[2] Davenport T, and Kalakota R. The potential for artificial intelligence in healthcare. Future Healthcare Journal. 2019; 6: 94–98.

[3] Lopez-Campos G, et al. Digital Interventions and Their Unexpected Outcomes - Time for Digitalovigilance? Stud Health Technol Inform. 2024; 310: 479–483.

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

[5] Balendran A, et al. Algorithmovigilance, lessons from pharmacovigilance. *NPJ Digit Med*. 2024; 7:270.

[6] Turri V, and Dzombak R. Why We Need to Know More: Exploring the State of AI Incident Documentation Practices. AIES, 2023: pp. 576–583.

[7] Rodrigues R, et al. When Artificial Intelligence Fails: The Emerging Role of Incident Databases. *PGAF*. 2023; 8:17–28.

[8] Zhang L, et al. Empirical Research in Software Engineering — A Literature Survey. *J. Comput. Sci. Technol.* 2018; 33:876–899.

[9] Blei DM, et al. Latent dirichlet allocation. *Journal of Machine Learning Research*. 2003;3: 993–1022.

[10] Burema D et al. A sector-based approach to AI ethics: Understanding ethical issues of AI-related incidents within their sectoral context. AIES, 2023: pp. 705–714.

[11] May R, et al. SoK: How Artificial-Intelligence Incidents Can Jeopardize Safety and Security. ARES, 2024: pp. 1–12.

[12] Agarwal A, and Nene MJ. Addressing AI Risks in Critical Infrastructure: Formalising the AI Incident Reporting Process. CONECCT, 2024: pp. 1–6.

[13] Stevens K, et al. Exploring Topic Coherence over Many Models and Many Topics. EMNLP, 2012: pp. 952–961.