CONDUCTSCIENCE
ACADEMIC PUBLISHING

Authors: Dr Joshua M Faber, Dr Chad Teven, Mr Arturo Francisco Balaguer Townsend
Type: Original Research
Keywords: AI, Bioethics, ChatGPT, Machine Learning
Edited By: Shuhan He, MD
Received: 2024-04-23
Published: 2025-02-22

A Scoping Review of Health AI Controversies in the Grey Literature from 2013-2022

ABSTRACT

Objective

This study aimed to 1) quantify the number of controversies related to health artificial intelligence (AI) development and deployment over the past 10 years, and 2) categorize health AI controversies by theme.

Materials and Methods

This study is a scoping review. We queried GoogleNews for articles meeting pre-specified inclusion and exclusion criteria related to health AI controversies. A total of 7508 articles were queried, with 76 articles ultimately meeting criteria. Articles were quantitatively evaluated for timing and coded according to both the organization and the themes of the controversies described in the article.

Results

Of the 76 articles, 55% were published in 2019 and 2022. No articles were published before 2016. Privacy, AI accuracy, and AI bias were the most common themes. Oversight and conflict of interest were the least common themes. Google, Babylon, and Optum were the most discussed companies.

Conclusions

The results demonstrate concrete evidence for many of the theoretical concerns highlighted by patient and physician surveys as well as scholarly ethics research. However, the volume of articles in the 10-year study period was small. It is possible that this study was too narrow in scope, that health AI has actually not yielded a large volume of controversial events, or that news organizations have insufficiently investigated potential health AI controversies. Further research is required to develop a more complete understanding of controversial activities in health AI.

BACKGROUND AND SIGNIFICANCE

AI has the potential to revolutionize medicine. Health AI tools are already being deployed to help providers diagnose more accurately, strengthen care plans, and alleviate provider burnout.[1][2] For these reasons, funding for AI in healthcare has increased by almost 100-fold over 10 years.[3]

Yet, despite such investment, health care providers and patients remain skeptical of its deployment to improve clinical care. One survey evaluating physician attitudes towards AI found that 49% of American physicians are uncomfortable with AI, and 20% of physicians believe AI has changed the way they practice medicine.[4] Another survey found that 80% of healthcare professionals expressed worry over how AI may compromise stakeholder privacy specifically, and 40% believed AI could be “more dangerous” than nuclear weapons.[5] One international integrative review found that physicians and nurses are concerned health AI may lead to less professional autonomy, increased legal uncertainty, underdevelopment of clinical skills, and job instability.[6]

Additionally, while patients are optimistic that AI may improve clinical care, they are also concerned about its clinical reliability. In a survey of 926 patients, 70% expressed discomfort with diagnoses made using an AI-based algorithm that was unable to further explain its reasoning beyond the initial diagnosis.[7] This “black box” or “explainability” problem is a long-standing concern in AI, both within and outside healthcare contexts. Patients also highlighted other concerns, including misdiagnosis, privacy breaches, less time with clinicians and safety.[7][8]

Ethicists have also expressed concern regarding AI in healthcare. One mapping review categorizes all ethical concerns regarding the use of AI in healthcare into one of three categories: a) epistemic, or relating to the intrinsic unreliability of the AI tools’ process and/or recommendations, b) normative, or relating to all the positive or negative impact health AI tools may have on stakeholders, or c) traceability, or the fact that unexplainable algorithmic errors by AI tools lead to uncertainty when trying to assign responsibility for clinical mistakes. [9]

To some extent, patients', physicians', and ethicists' concerns over how AI may be used in healthcare have been shaped by a few controversial cases involving commercial players. Three cases have been particularly influential over the past five years. In 2018, The New York Times and ProPublica published an investigative report on the relationship between Paige.AI, an AI-driven pathology company, and Memorial Sloan Kettering Cancer Center (MSKCC).[10] Leaders at MSKCC had invested in Paige.AI while simultaneously sharing data with the startup. Hospital physicians and scientists suggested a conflict of interest. They voiced concerns that hospital leadership had both invested in Paige.AI and shared valuable clinical resources with the company. The arrangement could be perceived as an attempt by executives to personally enrich themselves by sharing hospital property.

In 2019, multiple news outlets covered the *Dinerstein v. Google* lawsuit.[11] This case underscored AI-related privacy concerns after Dinerstein alleged that the University of Chicago shared his healthcare data without consent in order to support Google's development of an AI-based electronic health record.[12]

Thirdly, also in 2019, a study published in *Science* identified racial bias by one of Optum's widely utilized patient risk stratification algorithms. This bias led to an underinvestment of care coordination resources in non-white communities.[13]

It is unclear whether the MSKCC, Google, and Optum controversies constitute a handful of edge cases, or represent a broader pattern of corporate behavior in the AI healthcare space. If these cases are unrepresentative of the majority, ethicists, physicians, and patients may be wise to reconsider their concerns. However, if these cases are part of a broader pattern of ethically dubious corporate behavior, stakeholders would be fully justified in their concerns, including regulators who may be obligated to more rapidly act to prevent future violations. The answer, however, remains unclear. To that end, this scoping review takes an initial step towards evaluating stakeholder concerns by consolidating, categorizing, and summarizing health AI controversies documented in the news media over a ten year period.

The primary goals of the study are twofold: 1) to quantify trends in the number of health AI controversies from 2013-2022 and 2) to categorize health AI controversies by organization and theme

METHODS

Our protocol was modeled based on an approach to scoping reviews detailed previously by Arksey and O'Malley.[14]

We defined AI controversies in the following manner: Any event in which the data acquisition, training, or implementation of a health AI tool raises ethical or legal concerns resulting in a news article response. Ethical concerns include objections related to beneficence (obligations to act in patients' best interest), nonmaleficence (obligations to minimize patient harm), autonomy (obligations to respect patient preferences and decision-making), and justice (obligations to equitably serve patients). Legal concerns include legal objections to a health AI tool, such as lawsuits and publicized reviews by regulatory bodies.

Search terms were selected in collaboration with a librarian to identify AI controversies broadly, as well as controversies in the more specific AI fields of natural language processing (NLP) and machine learning (ML). A list of 43 search terms was eventually reduced to the 9 terms in Table 1 by removing duplicative terms, synonyms, and overly specific terms.

Table 1: Term Selection

Term 1	Term 2	Term 3
Health OR	Artificial Intelligence OR	Controversy OR
Medical OR	Machine Learning OR	Danger OR
Clinical	Natural Language Processing	Risk

Table 1 Term Selection

Search terms were queried in GoogleNews over a 10 year period from January 1, 2013 through January 1, 2023. Articles were sorted by relevance. Two reviewers evaluated up to 300 articles total in each query for inclusion in this review. Articles were initially included if their headlines met the following criteria:

1. Mentioned a specific commercial AI product use case or AI product organization
2. Used language suggesting illegal, unethical, or harmful behavior

The authors then read each article meeting inclusion criteria. Articles were subsequently excluded if they met the following criteria:

1. Did not focus on the healthcare industry
2. Focused on a controversy that is not AI related
3. Focused on academic research without a commercial sponsor or implication
4. Focused on a hypothetical or theoretical controversy instead of an observed controversy

Next, articles were coded across 3 domains (Table 2).

Table 2: Article Characteristics

Characteristics	Description
Theme	The nature of the controversies addressed in the article (e.g. privacy, conflict of interest, etc.)
Organization	The organization involved in the controversy
AI Controversy	Binary variable describing whether the training or deployment of the AI tool is the cause of the controversy outlined in the article

Table 2 Article Characteristics

The authors reviewed 300 of the same articles to test inter-reader reliability of inclusion criteria, exclusion criteria, and article labeling. Initial reliability was low, with Reviewer 1 including twice as many articles as Reviewer 2. After discussing the differences driving variation with a third reviewer, the inclusion and exclusion criteria were more explicitly defined to yield the above phrasing. As a result, reliability increased to nearly 100% on a subsequent sample of 300 articles. For the <5% of studies that led to different opinions between the initial reviewers, Reviewer 3 served to adjudicate any discrepancy.

Figure 1: Search Strategy

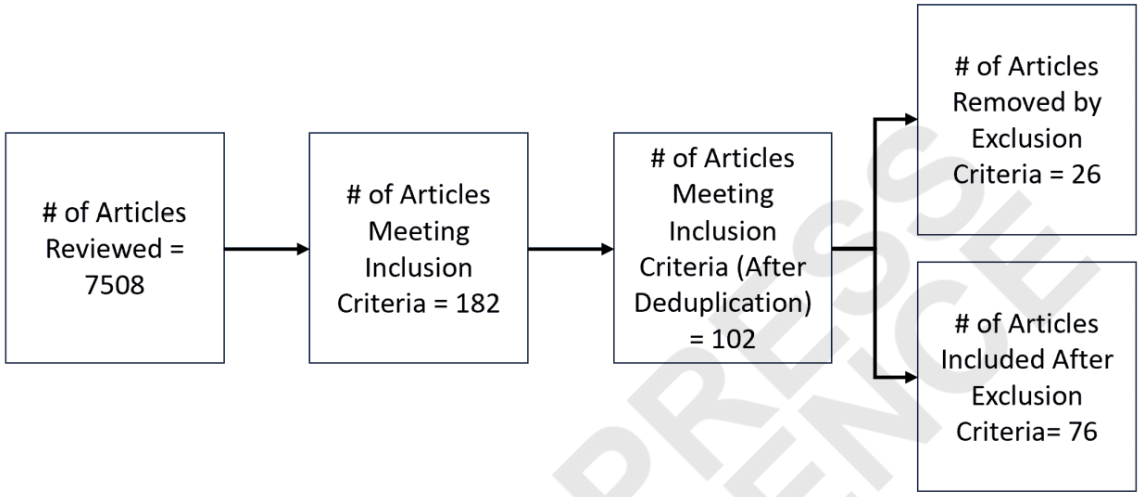

Figure 1 Search Strategy

Final results were generated based on 76 articles meeting inclusion criteria (Figure 1).

RESULTS

We identified 76 articles accessible via Google News discussing controversies in healthcare AI from 1/1/2013 to 1/1/2023. The articles reference 38 different organizations. Multiple articles detailed more than just one theme for more than just one organization. As a result, we identified 124 themes across our dataset.

Figure 2: Number of articles by year

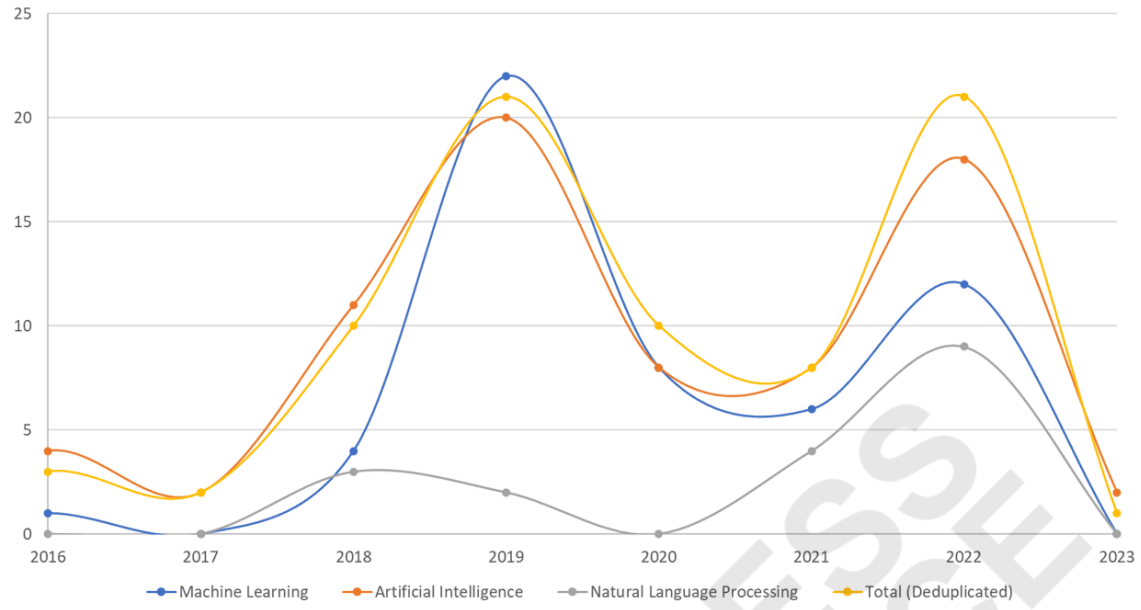

Figure 2 Number of articles by year

Article volumes fluctuated from year to year (Figure 2). There were no articles published before 2016 among the top 300 most relevant articles in each query. The 2019 spike in articles was largely driven by news media reactions to Google's Project Nightingale and demonstrated racial bias in Optum's patient risk stratification algorithm. The 2022 spike was driven by a broader array of health AI controversies, particularly relating to Loris and Google.

Queries related to ML and AI were far more common than articles related to NLP. Large language models (LLMs) were discussed in 16 articles, primarily related to OpenAI and Babylon. 1 article related to ChatGPT resulted from the query.

Table 3: Controversy Themes

Organization	# of Articles Highlighting a Given Theme												
	# of Articles	Socio-economic Bias	Poor Evidence	Business Concerns	Conflict of Interest	Exploitation	Lawsuit	Privacy	Oversight	Poor Accuracy	Poor Outcomes	Consent	Total Themes
Google	14		1	1			4	8	1		2	4	21
Babylon	9	1	4	2	1					3			12
Optum	9	9											9
NHS	8		2	1			2	4		3	1	2	14
Crisis text	5					3		2				1	6
Project Nightingale	5	1						5				1	7
Loris	4					3		1				1	5
OpenAI	3	1								1	1		3
Epic	3			1						2	1		4
Paige.AI	3				3			3					6
Royal Free London Trust	3						3						3
Facebook	2							2					2
Ascension	2							2				2	4
DeepMind	2			1			1	1					3
IBM	2									2			2
Nuance	2									2			2
Amazon	1							1					1
Rekognition	1							1					1
Exit International	1										1		1
InterAI	1										1		1
LAION-5B	1							1					1
M Health Fairview	1							1					1
Mayo Clinic	1							1					1
Moorfields	1							1					1
Northwell	1							1					1
Paraban Nanolabs	1							1					1
Sage Bionetworks DREAM Challenge	1	1									1		2
Sanford Health	1	1											1
South Korea	1							1					1
Stanford	1										1		1
Surgisphere	1		1										1
The Trevor Project	1	1											1
Watson	1									1			1
Woebot	1									1			1
Youper	1									1			1
Wysa	1									1			1
InterRAI	1	1											1
Total		**16**	8	6	4	6	10	**37**	1	**16**	9	11	124

Table 3 Controversy Themes

The top 3 most common themes (see Table 3) are privacy (37 instances), model accuracy (16 instances), and model socioeconomic (SE) bias (16 instances). The top 3 most commonly implicated companies are Google (19 articles), Babylon (9 articles), and Optum (9 articles). The British National Health Service (NHS) was the most commonly implicated non-commercial organization (8 articles).

Google's controversies centered mostly on privacy and consent concerns (48% of all Google-associated themes). Project Nightingale was the most common subject of controversy for Google. Babylon controversies were largely tied to concerns around poor evidence underpinning their AI (33% of all Babylon themes) and possible inaccuracies in the medical advice that their chatbot

would provide to patients (33% of all Babylon themes). A highly impactful *Science* article describing racial bias in one of Optum's population health tools drove news articles focused on the analytics company.[15] As a result, all Optum controversies were related to bias (100% of all Optum themes).

Relatively few articles highlighted conflict of interest concerns. The primary conflict of interest controversy over the past 10 years was Paige.AI's simultaneous use of Memorial Sloan Kettering (MSK) pathology data to build its core business and private investment in Paige.AI by individual members of MSK's board.[10]

Similarly, only one article highlighted data or AI oversight concerns. That article in Vox reported Google's decision to dissolve its AI ethics board.[16]

DISCUSSION

This study is meaningful because it represents the first scoping review of real AI controversies highlighted in the print news media. A large body of existing literature has used individual controversies as a jumping off point to discuss the theoretical ethical risks of AI in healthcare. A separate body of empirical research has leveraged clinician or patient opinion surveys. This study takes a first step towards substantiating theoretical concerns and stakeholder opinions through real world events. Research in this vein is now more essential than ever as governance bodies aim to regulate AI in healthcare. For instance, the World Health Organization's (WHO) 2024 guidance on large multi modal models cites over 150 academic and lay articles to justify an ethical framework for AI regulation. Yet, few of those articles highlight evidence of real events resulting from model training and deployment that involve illegal, unethical, or harmful outcomes. Instead, the WHO report relies primarily on theory and expert consensus to support its recommendations[17]. Future guidelines and regulations may be strengthened by an evidence base more explicitly rooted in real world events.

The first major result of this study is the low volume of real events reported by news organizations. We queried 10 years of GoogleNews articles to review 7508 articles, and only 76 met study criteria. Of those articles, 50% investigated events related to one of the top 5 most commonly referenced organizations in our dataset–Google, Babylon, Optum, the NHS, and Loris. Most of these organizations were investigated for just one or two controversial events, like Project Nightingale at Google or AI-based diagnostics at Babylon.

There are a number of possible reasons for this result. First, it is possible that the controversial use of health AI is actually much less common than suggested by theoretical studies and public opinion surveys. It may be that health AI organizations are very responsible in their data acquisition and AI deployment processes. On the other hand, it is also possible that news data is simply not a sensitive enough instrument to identify health AI controversies. Perhaps controversial behavior at health AI organizations is more common than these results would suggest and news media has not found it feasible or impactful to report on those controversial behaviors. This second hypothesis is substantiated by the fact that almost all organizations in our database are large entities with millions or billions of dollars in annual revenue. The skew in our data towards large organizations is inconsistent with the tens of billions of dollars invested in smaller, privately held health AI companies, especially startups, over the past several years.[18] It is likely that these small companies are more difficult to investigate than their larger peers. These companies have fewer employees who may serve as whistle blowers, are less likely to be subject to the shareholder scrutiny of publicly traded entities, and lack the type of massive user base that may uncover

harmful (and newsworthy) edge cases in their products. As a result, these companies may escape the kind of editorial investigation that would manifest as a data point in our queries.

The second major result of this study is that the most common themes for controversial events–privacy, accuracy, and bias—are also minimally regulated in the United States. Ironically, all three of these themes were recently highlighted by the World Health Organization (WHO) as key risks that ought to be addressed by governments as they regulate foundational models in AI.[17]

Privacy is the most strongly regulated topic via the Health Insurance Portability and Accountability Act (HIPAA), but much has changed in the field of digital health since its passage in 1996. At that time, the second “AI winter” had reduced AI funding significantly. The development of Watson by IBM would not begin for another 11 years.[19] The passage of HIPAA created a regulatory environment that is no longer up to date with modern advances in AI. Individual states have taken steps to build on HIPAA, but the federal government has been less active.[20] The relative abundance of privacy-related controversies revealed through this study suggest that individuals and news organizations either 1) believe health AI should be more closely regulated such that controversial events do not come to pass or 2) do not fully understand existing regulatory or ethical standards and, as a result, are surprised when organizations violate their perceived rights.

Regulations around AI accuracy and SE bias are nascent and underdeveloped. Recently, the Food and Drug Administration (FDA) and the Office of the National Coordinator (ONC) have taken the lead in approving and regulating AI algorithms for clinical use. FDA regulates algorithms that qualify as devices. FDA has defined devices as those algorithms that provide clinical decision support.[21] As of the date of this publication, FDA has not released regulatory guidance that explicitly specifies approaches to measure and optimize AI accuracy and SE bias. Moreover, FDA has not yet approved any devices that rely on generative AI.[22] In January 2024, ONC released a final rule pertaining to transparency in decision support intervention (DSI) development and deployment that focuses primarily on accuracy and bias, but also touches on privacy. Per ONC, decision support intervention developers must analyze their algorithms or models for “validity, reliability, robustness, fairness, intelligibility, safety, security, and privacy”. They must also have governance and mitigation strategies in place for data acquisition, management, and use. To this end, new certification will be required for developers beginning in 2025.[23] This final rule is a firm step in the right direction based on the results of this study.

The third major result of this study is the meaningful volume of LLM-related controversies, captured by our search terms. LLMs may be categorized as AI, ML, or NLP tools, and so all of our search terms had the potential to yield results related to LLMs. The majority of the 16 articles mentioning LLMs were related to chatbots. Two articles focused on Nuance’s note transcription services. Most articles discussing LLMs highlighted controversies related to poor model accuracy, outcomes, or evidence.

This study has limitations. First, this study is not a systematic review. As a result, the scope of this study is limited and the potential for bias in this study is elevated compared to a systematic review. Indeed, no formal assessment of bias of our underlying data was conducted. The scope of this study is also limited by the selected query terms, query size (maximum 300 articles per term), and database selection. Utilization of GoogleNews as the only queried database likely limited the number of articles captured by our study. It is also possible that the GoogleNews “relevance” filter that was applied to all of our queries filtered out articles that were truly relevant to the study but excluded on the basis of Google’s algorithm. Finally, this study only includes data through January 1, 2023. The enormous boom in the usage of AI and LLMs since 2022 has produced a flood of new grey literature that is not included in this study.

CONCLUSIONS

Our results begin to substantiate health AI ethical concerns espoused by theorists and measured through public opinion surveys. LLMs in particular have driven a meaningful amount of controversy over the past 8 years. Concerns related to privacy, model accuracy, and SE bias in AI are real and have manifested through multiple well-documented controversies. Limited federal regulation and limited public understanding of AI law and ethics may contribute to the occurrence of these controversies. That said, the number of articles captured by this study is lower than expected. It is possible that this study was too narrow in scope, that health AI has actually not yielded a large volume of controversial events, or that news organizations have insufficiently investigated potential health AI controversies. Thus, further research is required beyond this scoping review to identify the importance of each of these factors in explaining the relatively low volume of controversies identified in this study.

ACKNOWLEDGEMENTS

CONFLICT OF INTEREST STATEMENT

No conflicts of interest to report.

FUNDING

This research study was funded by the Northwestern Center for Bioethics and Humanities.

DATA AVAILABILITY

The data underlying this study will be shared upon reasonable request by the corresponding author.

REFERENCES

1. Briganti G and Le Moine O, Artificial Intelligence in Medicine: Today and Tomorrow. *Front. Med.* 2020;7:27.
2. Yeasmin S. Benefits of Artificial Intelligence in Medicine. In: 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS). 2019;1–6.
3. Ben Leonard, Ruth Reader. POLITICO. 2022 [cited 2023 Nov 8]. Artificial intelligence was supposed to transform health care. It hasn't. Available from: https://www.politico.com/news/2022/08/15/artificial-intelligence-health-care-00051828
4. Heather Landi. Fierce Healthcare. 2019 [cited 2023 Nov 8]. Nearly half of U.S. doctors say they are anxious about using AI-powered software: survey. Available from: https://www.fiercehealthcare.com/practices/nearly-half-u-s-doctors-say-they-are-anxious-about-using-ai-powered-software-survey
5. Castagno S, Khalifa M. Perceptions of Artificial Intelligence Among Healthcare Staff: A Qualitative Survey Study. Frontiers in Artificial Intelligence [Internet]. 2020;3:1-7. Available from: https://www.frontiersin.org/articles/10.3389/frai.2020.578983
6. Lambert SI, Madi M, Sopka S, Lenes A, Stange H, Buszello CP, et al. An integrative review on the acceptance of artificial intelligence among healthcare professionals in hospitals. npj Digit Med. 2023 Jun 10;6(1):1–14.
7. Khullar D, Casalino LP, Qian Y, Lu Y, Krumholz HM, Aneja S. Perspectives of Patients About Artificial Intelligence in Health Care. JAMA Network Open. 2022 May 4;5(5):e2210309.

8. Richardson JP, Smith C, Curtis S, Watson S, Zhu X, Barry B, et al. Patient apprehensions about the use of artificial intelligence in healthcare. npj Digit Med. 2021 Sep 21;4(1):1–6.
9. Morley J, Machado CCV, Burr C, Cowls J, Joshi I, Taddeo M, et al. The ethics of AI in health care: A mapping review. Social Science & Medicine. 2020 Sep 1;260:113172.
10. Ornstein C, Thomas K. Sloan Kettering's Cozy Deal With Start-Up Ignites a New Uproar. The New York Times [Internet]. 2018 Sep 20 [cited 2023 Nov 8]; Available from: https://www.nytimes.com/2018/09/20/health/memorial-sloan-kettering-cancer-paige-ai.html
11. Daisuke Wakabayashi. Google and the University of Chicago Are Sued Over Data Sharing. The New York Times [Internet]. 2019 Jun 26 [cited 2023 Nov 8]; Available from: https://www.nytimes.com/2019/06/26/technology/google-university-chicago-data-sharing-lawsuit.html
12. Lisa Schencker. Chicago Tribune. 2020 [cited 2023 Nov 8]. Judge dismisses lawsuit alleging University of Chicago Medical Center gave Google patient records without consent. Available from: https://www.chicagotribune.com/business/ct-biz-university-of-chicago-google-lawsuit-dismissed-patient-privacy-20200909-lbokttsv6rdc5aeen2q37naady-story.html
13. Quinn Gawronski. NBC News. 2019 [cited 2023 Oct 25]. Racial bias found in widely used health care algorithm. Available from: https://www.nbcnews.com/news/nbcblk/racial-bias-found-widely-used-health-care-algorithm-n1076436
14. Arksey H, O'Malley L. Scoping Studies: Towards a Methodological Framework. International Journal of Social Research Methodology: Theory & Practice. 2005;8(1):19–32.
15. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019 Oct 25;366(6464):447–53.
16. Kelsey Piper. Vox. [cited 2023 Nov 8]. Exclusive: Google cancels AI ethics board in response to outcry. Available from: https://www.vox.com/future-perfect/2019/4/4/18295933/google-cancels-ai-ethics-board
17. Ethics and governance of artificial intelligence for health. Guidance on large multi-modal models. Geneva: World Health Organization; 2024. Licence: CC BY-NC-SA 3.0 IGO.
18. Shah M. Generative AI shines amid VC funding chill [Internet]. Medium. 2023 [cited 2023 Nov 8]. Available from: https://medium.com/@Munjal-Shah/generative-ai-shines-amid-vc-funding-chill-51a960b70c45
19. Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. Gastrointestinal Endoscopy. 2020 Oct 1;92(4):807–12.
20. Michelson KN, Adams JG, Faber JMM. Navigating Clinical and Business Ethics While Sharing Patient Data. JAMA. 2022 Mar 15;327(11):1025–6.
21. US Food and Drug Administration. Clinical Decision Support Software - Guidance for Industry and Food and Drug Administration Staff. 2022 Sep 28:1-26.
22. Health C for D and R. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. FDA [Internet]. 2023 Oct 20 [cited 2023 Nov 8]; Available from:https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices
23. Office of the National Coordinator for Health Information Technology. Health data, technology, and interoperability: certification program updates, algorithm transparency, and information sharing. Fed Regist. 2022;87(48):13862-13925; 1197.