Discover Artificial Intelligence

Research

The role of explainability in AI-supported medical decision-making

Anne Gerdes[1]

Received: 29 August 2023 / Accepted: 21 March 2024
Published online: 29 April 2024

Abstract
This article positions explainability as an enabler of ethically justified medical decision-making by emphasizing the combination of pragmatically useful explanations and comprehensive validation of AI decision-support systems in real-life clinical settings. In this setting, post hoc medical explainability is defined as practical yet non-exhaustive explanations that facilitate shared decision-making between a physician and a patient in a specific clinical context. However, giving precedence to an explanation-centric approach over a validation-centric one in the domain of AI decision-support systems, it is still pivotal to recognize the inherent tension between the eagerness to deploy AI in healthcare and the necessity for thorough, time-consuming external and prospective validation of AI. Consequently, in clinical decision-making, integrating a retrospectively analyzed and prospectively validated AI system, along with post hoc explanations, can facilitate the explanatory needs of physicians and patients in the context of medical decision-making supported by AI.

1 Introduction

Within healthcare, deep learning artificial intelligence (AI) has proven good performance across different clinical decision support tasks, such as, e.g., improvement of mammography screening accuracy and reduction of workload. Lång et al. [1] have conducted the first randomized trial assessing the safety of an AI-supported screen reading procedure, including triage and detection of breast cancer. The authors conclude that the "AI supported screen-reading procedure enabled a 44·3% reduction in the screen-reading workload. The results indicate that the proposed screening strategy is safe" [1]. In addition, a recent retrospective population-wide mammography screening accuracy study concludes that "an AI system with an appropriate threshold could be feasible as a replacement of the first reader in double reading with arbitration." [2].

Progress in AI presents significant potential for positively impacting healthcare; simultaneously, it highlights the recognized importance of transparency in ensuring trustworthy AI, which is widely reflected in research, legislation, and public opinion. Consequently, explainable AI is viewed as a potential remedy for mitigating algorithmic opacity, playing a pivotal role in enhancing AI-supported decision-making [3–7]. Hence, concerning the use of automated decision-making and profiling, GDPR establishes the right for individuals to "meaningful information about the logic involved as well as the significance and the envisaged consequences of such processing for the data subject" (Article 13(f)), [8]. The European AI Act increases transparency obligations, implying that "High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately" (Article 13) [9]. Additionally, a Danish survey on peoples' preferences for AI reveals that the public's main priority is that physicians are responsible for the treatment. Second, the public demands explainability, and thirdly, that AI systems are tested for discrimination [10]. These findings have recently been confirmed in a study on people's

✉ Anne Gerdes, gerdes@sdu.dk | [1]University of Southern Denmark, Kolding, Denmark.

perceptions of AI in selected public (including medical diagnostics) and commercial sectors, which also reveals that "people want to know that AI has been used, and they want the human decision-maker to make the final decision" [11].

Given this backdrop, the article emphasizes that a combination of adequate retrospective evaluation of an AI system's performance and prospective validation of the AI system in a real-life clinical setting, along with post hoc explanations, can facilitate the explanatory needs of physicians and patients. Furthermore, post hoc medical explainability can be defined as reflecting non-exhaustive yet pragmatically useful explanations within a given clinical context.

The article is organized as follows: Section two clarifies the concepts of explainability and interpretability in AI. Against this backdrop, section three discusses selected approaches for improving transparency in AI-supported healthcare and elaborates on the notion of medical explainability. The focus is on addressing the patient's and the physician's needs for explanations by prioritizing the delivery of adequate information in an accessible manner, fostering comprehension, and cultivating trust in AI decision-making processes. Finally, section four concludes that a combination of proven AI system accuracy and post hoc explanations can facilitate the explanatory needs of physicians and patients.

2 The notion of explainable AI

Ensuring explainability and interpretability is crucial for bolstering transparency and trustworthiness in AI decision-making. The ability to explain decision-making processes comprehensibly and inspect and audit a model's inner workings implies that an AI system can be better understood, validated, and trusted by involved stakeholders and users, e.g., patients and clinicians. Yet, there is a notable absence of consensus on the precise definition, implying that explainability and interpretability are sometimes used interchangeably [12]. However, while this warrants a comprehensive examination, this article will not extensively explore the overall discussion surrounding this lack of agreement on definitions. Instead, and hopefully without further confusing matters, in the context of explainable AI, in this article, explainability is taken to refer to whether the AI system's decision-making can be explained in a manner understandable to humans. This can be obtained by developing an explanatory model approximating a black-box model's prediction outputs. As such, an explanatory algorithm does not make predictions; it presents a white box model that "partially mimics the behavior of the black box," and it is used for a post hoc explanation that accounts for the original model's prediction output [13]. In that sense, "explanations must be wrong" because the white box does not perfectly fit the black box; if that were the case, the two models would be identical [14].

Black box explainers, often referred to as model-agnostic methods, can provide textual or visual explanations of outputs. These approaches may, for example, showcase visual representations of feature weights, elucidate the significance of pixels in the context of image recognition (e.g., a heat map of an X-Ray), or offer insights into how modifications to input data can impact the final results (e.g., counterfactual explanations) [6].

In contrast, an AI system is said to be interpretable to the extent that one can scrutinize its inner workings, i.e., analyze the model's structure, parameters, and features to clarify how an AI system arrives at a decision. A machine learning model might be inherently explainable (white box or glass box model) because it reflects a simple relationship between input and output data, which is scrutable and straightforwardly understandable (e.g., linear regression models and decision trees). Conversely, complex black-box models require creating a post hoc model to be understandable to end users. Furthermore, there is a widespread assumption that the most complex machine learning models perform best [3, 4, 10, 15]. However, it is worth noting that the connection between the complexity of a machine learning model and its accuracy is not straightforward. While a simple model can excel on specific data, it might fall short when faced with other data or more intricate problems. One could argue, in line with Occam's razor, that if two models achieve the same level of accuracy for a particular problem, it is preferable to opt for the simpler one. In essence, the model should strive to be as simple as possible while possessing the necessary complexity to tackle the given task effectively. Rudin notes that "[t]he belief that there is always a tradeoff between accuracy and interpretability has led many researchers to forget the attempt to produce an interpretable model" [14]. Therefore, interpretable models have not gained traction,[1] and she suggests that researchers devote more resources to working on inherently explainable models rather than working on the post hoc explainability of deep neural networks.

[1] According to Rudin [14], we lack analysts with the skills to create interpretable models.

3 What should we demand of medical explainability?

Although we may not be able to completely unveil the workings of an AI system, physicians and patients must be helped to understand the suitability of AI-generated input to decision-making in a specific clinical setting. Hence, Lipton [16] introduces the notion of contestability, implying that an explanation must present the reasoning behind a decision and offer an opportunity to contest and modify a wrong decision. This notion of contestability is further elaborated on by Ploug and Holm [17]. They acknowledge that many facets of human and diagnostic reasoning remain elusive to complete explanation. Therefore, every aspect of AI decision-making does not necessarily have to be fully explainable. Instead, they suggest ensuring that patients have access to adequate information, empowering them to question AI decision-making effectively. Consequently, the authors address explainability challenges in healthcare by proposing a right for patients to contest AI diagnostic decisions, which, subsequently, ought to guide the notion of transparency in AI decision-making.

According to the authors, a right to contest can be demarcated as a right grounded in normative principles concerning what counts as morally relevant information when protecting patients' rights and protecting them against harm, including attention to what kind of claims to information patients reasonably can make when contesting AI medical decision-making. Furthermore, a contestability right must allow patients to effectively protect themselves by enabling them to contest a specific system by providing appropriate specificity instead of high-level information of little or no use (e.g., "this clinic use AI technology"). Moreover, this right must be proportional, implying that it is a right to self-protection while balancing others' right to benefit from access to AI healthcare. Notably, the authors emphasize that "the right to contest cannot justifiably entail transparency requirements—i.e., a right to information—that effectively makes impossible the very development and use of AI in health care" [17]. Finally, the right to contest AI decision-making aligns with the general right to contest medical decisions. In this context, the right to contestability is elaborated by introducing four domains for explanation: the context of health data, risk of bias, AI accuracy, and the role of AI in the decision-making process. Within these respective domains of contestable AI, i.e., data, bias, performance, and decision involvement, the authors specify "contestability variables" leading to contestability questions and required "contestability information" [17]. For example, gaining information about personal health data and data curation empowers patients by enabling them to contest an AI decision based on evaluating the sensitivity and quality of data. Likewise, information concerning bias mitigation, AI robustness, performance accuracy, and degree of human oversight is essential for patients to enforce a right to contest AI medicine decision-making. The authors rightly note that "if the right to contest should have any practical relevance for patients, the information must be made available in ways taking into account their ability to and interests in processing such information and by extension to the ability of healthcare professionals to process and communicate the information" [17].

The notion of contestability is essential and contributes valuable insights to inform the assessment of AI decision-making in healthcare and to demarcate patients' rights. Regrettably, the proposed contestability framework seems to be tailored to an audience familiar with AI health care issues. Consequently, while the authors acknowledge the importance of presenting contestability information in a patient-friendly manner, their framework doesn't necessarily cater to the needs of patients. Even digitally savvy patients knowledgeable about, e.g., data science, might not be able to judge whether clinical data sets have been curated adequately because they lack domain expertise. There might be nuances to this argument. Richardson et al. [18] report findings from 15 focus group interviews with 87 participants (18–91 years) investigating peoples' beliefs and attitudes toward the use of AI in healthcare. Asked to reflect on vignettes describing AI use cases in healthcare, people referred to experiences with AI narratives and established non-medical technologies, i.e., smartphones. Moreover, people recognized issues related to data ethics, i.e., AI systems trained on biased data may preserve existing forms of bias in healthcare and reinforce "stigma or limiting access" [18]. While this study doesn't specifically explore (or reveal findings concerning) the aspect of explainability, its findings suggest that some digitally literate patients might be able to navigate a contestability framework such as the one presented by Ploug and Holm [17]. Nonetheless, patients might also be overwhelmed and cognitively overloaded even when information is presented in a user-friendly format because this type of information comes on top of diagnostically relevant information concerning their health condition. In addition, when faced with the information needed to make informed decisions, patient autonomy does not equal rational self-determination; rather, it should be viewed in the context of interdependence as relational autonomy, emphasizing that the patient's circumstances, relationships, and social background are crucial factors in shaping their self-identity and capabilities when engaging in contexts of shared decision-making [19].

Still, the contestability framework has an excellent fit at the level of guidelines in healthcare for standardized and comprehensive reporting related to the assessment of AI, such as e.g., the CLAIM guideline checklist for AI in medical imaging [20], which facilitates clinical researchers in preparing and reviewing research papers by providing a list consisting of forty items addressing key concerns related to aspects concerning, e.g., data processing and curation and model development and validation. Also, Lekadir et al. [21] propose an "AI passport' for standardized description and traceability of medical AI tools." Similarly, the MAS-AI model [22], designed for evaluating AI in medical imaging, aims to assist decision-makers, such as medical directors, heads of hospital departments, local or national treatment councils, and procurement organizations. By offering guidance in adoption decisions, the MAS-AI model streamlines the process of determining the suitability of AI in healthcare settings. Furthermore, the reporting and validation guideline CONSORT (Consolidated Standard of Reporting Trials) has included AI in the CONSORT-AI statement, drawing attention to AI-specific items that are important in reporting [23]. Hence, the contestability framework suggested by Ploug and Holm [17] can be seen as valuable for clinical, administrative, and technical domain experts assessing AI systems in healthcare. In line with, e.g., the MAS-AI assessment framework, Ploug and Holm [17] caution against solely focusing on algorithmic transparency and draw attention to a broader perspective that includes contextual factors in the development and deployment of AI in healthcare.

In light of Ploug and Holm's recommendation [17] to avoid exclusively concentrating on model transparency and considering AI systems in isolation, the following discussion advances arguments from the viewpoint of what Holm [24] refers to as an *explanation view*. In line with this, when a non-transparent AI decision-support system undergoes comprehensive retrospective analysis of its performance and prospective validation in a real-world clinical setting using the randomized clinical trial model, incorporating an approximate post hoc explanatory model is sufficient to establish legitimacy. This, in turn, reinforces the physician's responsibility for treatment and contributes to an informed, collaborative decision-making process between the physician and patient.

Conversely, the validation perspective prioritizes AI performance accuracy over explainability, acknowledging the inherent challenge of achieving complete transparency as explainable AI is still an open challenge. From an AI safety and reliability perspective, proponents of this perspective assert that, besides being useful for auditing models, explainable AI is not the proper remedy for clinically deployed models [13, 25, 26]. Additionally, this view points to similarities between the unexplainable and non-causal nature of established clinical care practices and the black-box nature of AI systems, noting that "the most powerful machine learning approaches are not radically different from routine aspects of medical decision-making" [27]. As such, this view advocates for prioritizing comprehensive model validation, claiming that "unless there are substantial advances in explainable AI, we must treat these systems as black boxes, justified in their use not by just-so rationalisations, but instead by their reliable and experimentally confirmed performance" [25].

On the other hand, an *explainability view* may argue that pointing to the acceptance of opacity of medical decisions could equally well underscore that relative explainability suffices in meeting the demand for transparency when paired with a certification of an AI system [28]. Acknowledging bounded rationality, AI should not be measured against an idealized or double standard [29] of diagnostics claiming full AI explainability since medical diagnostic explanations are inherently incomplete and inaccessible i.e., "the optimization process of diagnostic is never fully achieved (…). The expectations of physicians lie in making decisions that balance explainable reasoning, procedural speed, reasonable levels of accuracy, and potentially a minimization of bias" [28]. Consequently, making ethical decisions in healthcare relies on *phronesis* or practical wisdom [30], where knowledge and experience allow the physician to identify the epistemically justified and morally relevant aspects within concrete situations. Clearly, as the physician is responsible for an explanation and accountable for mistakes, whereas an AI is not, this suggests that we ought to demand a higher standard of AI explanations, which seems to speak in favor of the *validation view* because a complex deep neural network cannot be explained and therefore, we ought to focus on validated performance accuracy. However, "[h]aving some medical AI certified according to the acceptable levels of accuracy dissolves the need for it to be explainable in detail to patients to confirm the conditional good of explainability" [28]. As such, a post hoc explanation may facilitate the explanatory needs of the parties within a specific clinical context.

As an example thereof, striving to render explainability beneficial for both physicians and patients, Arbelaez Ossa et al. [31] propose focusing on the specific clinical context. Hence, a black-box model may be tolerable in terms of medical explainability if "its usability within a specific clinical context is understood" [31]. Consequently, aligning with the transparency practices prevalent in healthcare overall and within the specific context of the clinical implementation of AI, it is reasonable to propose a spectrum of explainability requirements. This suggests that an explanation need not be exhaustive to be considered epistemically justifiable in improving comprehension regarding how AI decision-making inputs influence diagnostic decisions. According to their approach, which shares similarities with

the contestability framework of Ploug and Holm [17] and the certifiability approach of Kempt et al. [28], explainability involves assessing context-specific aspects related to data quality and curation, utility, accuracy, and uncertainty in performance, mitigation of biases, and clarification of issues related to responsibility and accountability.

In contrast to highlighting contextual factors, the validation perspective addresses the opacity issue primarily from an isolated technological standpoint rather than considering an AI system as embedded within the broader healthcare context. However, it is difficult to imagine, and also well-known from the above-mentioned studies [10, 11], that physicians and patients would consider it legitimate if AI systems recommend a treatment solely based on the validation and reliability of the AI system [32]. Consequently, within a specific clinical context, a suitable post hoc explanation should be curated to facilitate a pragmatically helpful explanation of the AI decision-making rationale. For instance, Bjerring and Busch [32] present a scenario in which a non-transparent but validated diagnostically accurate AI risk prediction system ranks multiple treatment options for a breast cancer patient and recommends surgical removal of both breasts. Without an additional post hoc explanation of how the system arrived at the recommendation, the physician, in consultation with the patient, cannot responsibly decide on treatment without overruling essential principles behind patient-centered medicine. The clinical judgments made by the physician suffer from epistemic ambiguity and lack justifiability, rendering them unsuitable for dialoguing with the patient. As a result, the patient, lacking sufficient information, cannot give informed consent. Consequently, the prerequisites for achieving shared decision-making are unmet [32].

On the contrary, "lower levels of explainability are manageable in tasks that do not increase patients' morbidity or mortality risk" [31]. Hence, a lack of explainability may be acceptable when the stakes are low. As such, Zerilli et al. [29] emphasize that "[p]rima facie, standards of transparency ought to be sensitive to the *stakes* of a decision, not to *who* or *what* is making it." As an example of such a low-risk scenario, a black-box algorithm for scoring knee osteoarthritis implemented at Bispebjerg and Frederiksberg University Hospital is used to triage patients referred for an MRI scan by their general practitioner. The algorithm is trained to align with the Danish clinical Guidelines for Knee Osteoarthritis for decisions concerning who is referred to MRI scanning. The algorithm produces an image analysis and a knee osteoarthritis report. If the algorithm does not predict knee osteoarthritis, the patient is MRI scanned immediately after, whereas the patient is referred to further consultations in case of knee osteoarthritis. The black-box algorithm autonomously reaches decisions without human intervention. However, human oversight follows, scrutinizing the decision-making process and performance accuracy. Yet, the algorithm is not trained to respond to cases in which the patient's condition overrules the Danish clinical guidelines, e.g., if the patient has suffered a broken leg, it necessitates an MRI scan even though the patient has knee osteoarthritis. Therefore, to approve or override the algorithmic decision without finding themselves in a state of unacceptable epistemic ambiguity, the radiologist needs to have an overall understanding of how the algorithm has been trained and validated. Furthermore, they need to know that the algorithm's performance is being monitored and that it adheres to the Danish clinical guidelines and holds certification in line with testing and validation criteria. A patient with knee osteoarthritis might question why they are not referred for an MRI scanning. In that case, they also only need to know that the algorithm makes decisions in accordance with the Danish clinical guidelines, which do not approve an MRI scan for patients with knee osteoarthritis, and that the system's performance is certified and closely monitored. If interested, the patient can gain insight into the certification process, i.e. how the algorithm is tested and validated. Thus, in this low-stakes scenario, the patient and radiologist do not need an additional post hoc explanation of the algorithmic output.

Consequently, compared with the above-mentioned AI risk prediction system for breast cancer treatment, the algorithm for scoring knee osteoarthritis needs to satisfy fewer explainability criteria. However, by prioritizing an explanation-centric approach over a validation-centric one in the realm of AI decision-support systems, it is still crucial to acknowledge the inherent tension between the, often times, politically driven enthusiasm to deploy AI in healthcare and the imperative for thorough, time-consuming validation with attention to both retrospective evaluation as well as prospective validation of an AI system in the clinical site of implementation. Hence, high-performance accuracy on data sets used to develop an algorithm does not necessarily transfer to external data sets and sites [33, 34]. Therefore, an assessment of performance consistency requires meticulous and extended trial evaluations, including adopting pilot phases where AI decision-making inputs are assessed without immediate application. Within this comprehensive validation context, a robust foundation for post hoc medical explainability can be established, characterized by non-exhaustive yet pragmatically useful explanations tailored to specific clinical contexts. Consequently, in clinical decision-making, the convergence of retrospective evaluation and prospective on-site validation, coupled with post hoc explanations, can be viewed as supporting the explanatory needs of both physicians and patients.

4 Conclusion

This article emphasizes post hoc medical explainability as necessary for the ethically justified use of AI decision-support systems in healthcare. As such, post hoc explainability is defined as pragmatically useful but not necessarily exhaustive explanations that enhance shared decision-making between healthcare professionals and patients in specific clinical contexts. At the same time, the article underscores the importance of combining post hoc explanations with thorough validation of AI decision-support systems within real-life clinical settings.

Moreover, physicians bear the responsibility for explanations and are accountable for errors, whereas AI lacks this accountability. This observation seems to lean toward prioritizing validated performance accuracy instead of explainability, especially considering that explainable AI is still an open challenge. Nevertheless, and in synthesizing the views expressed in [17, 28, 29, 31], having an AI system certified according to acceptable accuracy levels can relax the claim of exhaustive explanations. At the same time, it is essential to recognize the tension between the enthusiasm to deploy AI in healthcare and the demand for thorough, time-consuming retrospective and prospective validation to ensure reliability. In light of these considerations, the article suggests that integrating a retrospectively analyzed and prospectively validated AI system, alongside post hoc explanations, can cater to the explanatory needs in clinical decision-making contexts.

Acknowledgements The author is grateful for the helpful feedback from RT, MSc, Janus Uhd Nybing, CTO, Co-Founder Radiological AI Test Center, Department of Radiology, Bispebjerg-Frederiksberg Hospital, University of Copenhagen, Denmark. The author would also like to thank Professor Arthur Zimek, Department of Mathematics and Computer Science, University of Southern Denmark, for valuable comments on the relation between model complexity and accuracy. Finally, the author would like to thank the reviewers for their thorough reviews and insightful remarks, which played a crucial role in shaping this article.

Author contributions Anne Gerdes is the sole writer of this paper. The author read and approved the final manuscript.

Funding Open access funding provided by University of Southern Denmark.

Data availability The author does not analyze or generate any datasets since the focus of this work is purely theoretical.

Declarations

Competing interests The authors declare no competing interests.

References

1. Lång K, et al. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. Lancet Oncol. 2023;24(8):936–44.
2. Elhakim MT, et al. Breast cancer detection accuracy of AI in an entire screening population: a retrospective, multicentre study. Cancer Imaging. 2023;23(1):127.
3. Gunning D, Aha DW. DARPA's explainable artificial intelligence (XAI) program. AI Mag. 2019;40(2):44–58.
4. Gunning D, et al. DARPA's explainable AI (XAI) program: a retrospective. Appl AI Lett. 2021;2(4):1-11.
5. Danaher J. The threat of algocracy: reality, resistance and accommodation. Philos Technol. 2016;29(3):245–68.
6. Goebel R, et al. Explainable AI: the new 42? In: Holzinger A, et al., editors. Lecture notes in computer science (including subseries Lecture notes in artificial intelligence and lecture notes in bioinformatics). Cham: Springer International Publishing; 2018. p. 295–303.
7. Ribeiro MT, Singh S, Guestrin C. Why should i trust you? Explaining the predictions of any classifier. In: KDD '16: the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016. San Francisco: Association for Computing Machinery. p. 1135–44.
8. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance) [Internet]. OJ L, 32016R0679 May 4, 2016. 2016.

9. Proposal for a regulation of the European Parliament and the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. EUR-Lex - 52021PC0206. 2021, Eupean Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELLAR:e0649735-a372-11eb-9585-01aa75ed71a1.
10. Ploug T, et al. Population preferences for performance and explainability of artificial intelligence in health care: choice-based conjoint survey. J Med Internet Res. 2021;23(12):e26611.
11. Holm S, Ploug T. Population preferences for AI system features across eight different decision-making contexts. PLoS ONE. 2023;18(12):1.
12. Goisauf M, Cano Abadía M. Ethics of AI in radiology: a review of ethical and societal implications. Front Big Data. 2022;5:850383.
13. Babic B, et al. Beware explanations from AI in health care. Science. 2021;373(6552):284–6.
14. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–15.
15. Gerlings J, Jensen MS, Shollo A. Explainable AI, but explainable to whom? An exploratory case study of xAI in healthcare. In: Lim CP, Chen YW, Vaidya A, Mahorka C, Jain LC. (eds) Handbook of Artificial Intelligence in Healthcare. Intelligent Systems Reference Library Springer Cham 2022;212.https://doi.org/10.1007/978-3-030-83620-7_7
16. Lipton ZC. The mythos of model interpretability. Commun ACM. 2018;61(10):36–43.
17. Ploug T, Holm S. Right to contest AI diagnostics: defining transparency and explainability requirements from a patient's perspective. In: Artificial intelligence in medicine. 2022, Springer. p. 227–38.
18. Richardson JP, et al. A framework for examining patient attitudes regarding applications of artificial intelligence in healthcare. Digital Health. 2022;8:205520762210890.
19. Shih P, et al. Relational autonomy in breast diseases care: a qualitative study of contextual and social conditions of patients' capacity for decision-making. BMC Health Serv Res. 2018;18(1):818.
20. Mongan J, Moy L, Kahn CE Jr. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell. 2020;2(2):e200029.
21. Lekadir K, et al. Artificial intelligence in healthcare: applications, risks, and ethical and societal impacts. 2022.
22. Fasterholdt I, et al. Model for assessing the value of artificial intelligence in medical imaging (MAS-AI). Int J Technol Assess Health Care. 2022;38(1):e74.
23. Liu X, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digit Health. 2020;2(10):e537–48.
24. Holm S. On the justified use of AI decision support in evidence-based medicine: validity, explainability, and responsibility. Cambr Q Healthc Ethics 2023: p. 1–7. https://doi.org/10.1017/S0963180123000294 Epub ahead of print. PMID:37293823.
25. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745–50.
26. Graham SS. The doctor and the algorithm: promise, peril, and the future of health AI. Oxford: Oxford University Press; 2022.
27. London AJ. Artificial intelligence and black-box medical decisions accuracy versus explainability. Hastings Cent Rep. 2019;49(1):15–21.
28. Kempt H, Heilinger J-C, Nagel SK. Relative explainability and double standards in medical decision-making. Ethics Inf Technol. 2022;24(2):20.
29. Zerilli J, et al. Transparency in algorithmic and human decision-making: is there a double standard? Philos Technol. 2019;32(4):661–83.
30. Rackham H. Aristotle, The Athenian Constitution, The Eudemian ethics, on virtues and vices. With an English translation. Cambridge: Harvard University Press; 1935.
31. Arbelaez Ossa L, et al. Re-focusing explainability in medicine. Digit Health. 2022;8:20552076221074490.
32. Bjerring JC, Busch J. Artificial intelligence and patient-centered decision-making. Philos Technol. 2021;34(2):349–71.
33. Yu AC, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. Radiol Artif Intell. 2022;4(3):e210064.
34. Wang X, et al. Inconsistent performance of deep learning models on mammogram classification. J Am Coll Radiol. 2020;17(6):796–803.