INSIGHTS-JOURNAL OF

LIFE AND SOCIAL SCIENCES

INSIGHTS-JLSS

Insights-Journal of Life and Social Sciences

EVALUATING THE APPLICATION OF MACHINE LEARNING ALGORITHMS IN PREDICTING DISEASE OUTCOMES AND ENHANCING DIAGNOSTIC ACCURACY IN HEALTHCARE SYSTEMS

Systematic Review

Amjad Shakeel[1]*, Munib Sultan[2], Syed Hassan Idrees[3], Imran Nazir[4], Iqra Afzaal[5], Muddasir Ali[6], Bilal Yousaf[7]

[1]QA Executive, Pakistan Pharmaceutical Company Pvt. Ltd., Karachi, Pakistan.

[2]Liaquat National Hospital and Medical College, Karachi, Pakistan.

[3]MBBS, DPH, DOMS, MS Scholar (Healthcare Management), Riphah International University, Islamabad, Pakistan. Postgraduate Resident, Department of Ophthalmology, Al-Shifa Trust Eye Hospital, Rawalpindi, Pakistan.

[4]Student, BS Intelligent Systems and Robotics, Department of Information and Communication Engineering, The Islamia University of Bahawalpur, Bahawalpur, Pakistan.

[5]Rawalpindi Women University, Rawalpindi, Pakistan.

[6]Student, Department of Biochemistry, Gulab Devi Educational Complex, Lahore, Pakistan.

[7]Student, University of Okara, Okara, Pakistan.

Corresponding Author: Amjad Shakeel, QA Executive, Pakistan Pharmaceutical Company Pvt. Ltd., Karachi, Pakistan, amjad15shakeel@gmail.com

Conflict of Interest: None

Grant Support & Financial Support: None

Acknowledgment: The authors extend their sincere gratitude to the librarians for their invaluable assistance in developing and executing the comprehensive search strategy. We also thank our colleagues for their insightful feedback during the preparation of this manuscript. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

ABSTRACT

Background: The integration of machine learning (ML) into healthcare systems offers a transformative potential for enhancing disease prediction and diagnostic accuracy. Despite a surge in primary research, a comprehensive synthesis of the current evidence on the efficacy of ML algorithms across diverse clinical settings is lacking.

Objective: This systematic review aims to evaluate the application of machine learning algorithms in predicting disease outcomes and improving diagnostic accuracy within modern healthcare systems.

Methods: A systematic review was conducted following PRISMA guidelines. Electronic databases (PubMed, Scopus, Web of Science, Cochrane Library) were searched for studies published between 2019 and 2024. Included studies evaluated ML models for diagnosis or prognosis in human patients compared to standard care. Two reviewers independently screened studies, extracted data, and assessed risk of bias using appropriate tools (QUADAS-2, ROBINS-I).

Results: From 2,847 initial records, 8 studies were included. The studies encompassed oncology, ophthalmology, and cardiology, utilizing data from medical imaging and electronic health records. ML models, particularly deep learning algorithms, demonstrated high performance, frequently matching or surpassing clinical expert accuracy. Key metrics included area under the receiver operating characteristic curve (AUC-ROC) values often exceeding 0.90. Common limitations included retrospective design and risks to generalizability.

Conclusion: Machine learning algorithms show significant promise in enhancing diagnostic and prognostic precision, potentially supporting clinical decision-making. However, the current evidence is primarily derived from retrospective studies. Future research requires robust prospective validation and standardized reporting to ensure reliability and facilitate successful clinical integration.

Keywords: Machine Learning, Artificial Intelligence, Diagnosis, Prognosis, Systematic Review, Healthcare.

INTRODUCTION

The increasing complexity of modern healthcare, coupled with the rising global burden of both communicable and non-communicable diseases, has intensified the need for advanced analytical tools to support clinical decision-making. In this context, the integration of artificial intelligence, particularly machine learning (ML), into healthcare systems represents a paradigm shift with the potential to revolutionize patient care. Machine learning algorithms excel at identifying complex, non-linear patterns within large, high-dimensional datasets, a capability that is increasingly being harnessed to predict disease outcomes and enhance diagnostic accuracy (1). The clinical significance of this application is profound, as timely and accurate prediction can facilitate early intervention, personalize treatment strategies, optimize resource allocation, and ultimately improve patient prognoses while reducing healthcare costs. Current healthcare paradigms are often reactive, with interventions typically initiated after the manifestation of clear clinical symptoms. This approach is challenged by epidemiological realities; for instance, cardiovascular diseases remain the leading cause of death globally, accounting for an estimated 17.9 million lives annually, while cancer burden is projected to rise to 28.4 million cases by 2040 (2, 3). Early and accurate prediction is critical in mitigating these statistics. Although numerous studies have explored individual ML models for specific conditions—such as random forests for diabetes prediction or convolutional neural networks for radiological image analysis—the existing knowledge is fragmented. A significant gap in the literature is the lack of a comprehensive synthesis that critically evaluates and compares the efficacy of diverse ML algorithms across a spectrum of diseases and healthcare settings. Furthermore, the rapid evolution of these technologies necessitates a contemporary analysis of their real-world clinical validity and utility, justifying the need for a systematic review to consolidate and appraise the latest evidence (4).

The primary research question guiding this systematic review, structured using the PICO framework, is: In patients undergoing diagnostic evaluation for various diseases within modern healthcare systems (Population), how does the application of machine learning algorithms (Intervention) compare to standard diagnostic or predictive methods (Comparison) in terms of improving predictive accuracy for disease outcomes and enhancing diagnostic precision (Outcomes)? The objective is to systematically review and analyze the current evidence on the efficacy and accuracy of data-driven ML applications in clinical prediction and diagnosis. To address this question comprehensively, this systematic review will consider primary research studies, including randomized controlled trials, prospective and retrospective cohort studies, and diagnostic accuracy studies, that evaluate supervised and unsupervised ML models. The scope will be global, including studies published in the last five years (2019-2024) to ensure the findings reflect the most current advancements in this rapidly progressing field. This timeframe is critical as it captures the period of most significant innovation in deep learning and the expansion of ML into novel clinical domains. By adhering to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, this review aims to provide a rigorous and transparent synthesis of the evidence (5). The expected contribution of this work is to provide clinicians, healthcare administrators, and policy-makers with an evidence-based overview of the practical value and performance of various ML algorithms. It will highlight the most promising models for specific clinical applications, discuss common challenges related to data quality, model interpretability, and integration into clinical workflows, and identify key areas for future research. By critically evaluating the transition of ML from theoretical promise to clinical application, this review will serve as a foundational reference to guide the responsible and effective implementation of data-driven approaches in healthcare, thereby moving towards a more proactive and personalized model of medicine.

METHOD

This systematic review was conducted in strict accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to ensure methodological rigor and reproducibility (5). A comprehensive and systematic search strategy was designed and executed to identify all relevant published literature examining the application of machine learning (ML) algorithms for disease prediction and diagnostic enhancement. The electronic bibliographic databases searched included PubMed/MEDLINE, Scopus, Web of Science Core Collection, and the Cochrane Central Register of Controlled Trials. The search strategy incorporated a combination of controlled vocabulary terms, such as MeSH in PubMed, and free-text keywords related to the core concepts: "machine learning," "artificial intelligence," "deep learning," "diagnosis," "early diagnosis," "disease progression," and "prognosis." These terms were combined using appropriate Boolean operators (AND, OR) and tailored to the syntax requirements of each respective database. To mitigate the risk of omitting pertinent studies, the reference lists of all included articles and relevant review papers were manually screened. The study selection process was governed by pre-defined inclusion and exclusion criteria. Studies were eligible for inclusion if they were primary research articles published in English between January 2019 and April 2024, which evaluated a machine learning model for predicting a disease outcome or enhancing diagnostic accuracy in a human patient population. The intervention of interest was the application of any supervised, unsupervised, or deep learning algorithm (e.g., random forest, support vector machines, neural networks). The comparison was standard clinical diagnostic or prognostic methods. Measured outcomes included, but were not limited to, metrics of diagnostic performance such as accuracy, sensitivity, specificity, area under the receiver operating characteristic curve (AUC-ROC), or prognostic precision.

All study designs were considered, including randomized controlled trials, cohort studies, case-control studies, and diagnostic accuracy studies. Exclusion criteria encompassed review articles, editorials, conference abstracts without full text, studies not published in English, in vitro or animal studies, and studies where the ML model was not directly applied to a clinical diagnostic or prognostic task. The identification and selection of studies were performed by two independent reviewers to minimize selection bias. All retrieved citations were imported into the reference management software EndNote X20, where duplicates were removed. The subsequent screening process was managed using the Rayyan online systematic review platform (6). The initial screening was based on titles and abstracts, after which the full texts of potentially relevant articles were retrieved and assessed in detail against the eligibility criteria. Any disagreements between the reviewers at any stage of the selection process were resolved through discussion or, if necessary, by consultation with a third reviewer. This process was documented using a PRISMA flow diagram, which detailed the number of records identified, included, and excluded at each stage, along with the specific reasons for exclusions. Data from the included studies were extracted into a standardized, pilot-tested data extraction form to ensure consistency. The extracted variables included first author, publication year, country of origin, study design, target disease condition, patient population characteristics and sample size, type of data used (e.g., imaging, genomic, electronic health record data), specific machine learning algorithm(s) implemented, the comparator group (if any), and the primary outcome measures with their respective results.

The risk of bias and quality of the included studies were critically appraised using appropriate tools tailored to the study design. For randomized controlled trials, the Cochrane Risk of Bias 2 (RoB 2) tool was employed (7). For non-randomized studies, including cohort and case-control studies, the Risk Of Bias In Non-randomized Studies - of Interventions (ROBINS-I) tool was utilized (8). The quality of diagnostic accuracy studies was assessed using the QUADAS-2 tool (9). This assessment was also conducted by two independent reviewers. Given the anticipated heterogeneity in the target diseases, machine learning methodologies, data types, and outcome measures across the included studies, a quantitative meta-analysis was deemed unfeasible. Therefore, the data synthesis will be primarily qualitative, employing a narrative synthesis approach. The findings will be structured and summarized to describe the landscape of ML applications in healthcare, highlighting the types of algorithms that demonstrate superior performance for specific clinical tasks, the data modalities that yield the most robust models, and the common challenges and limitations reported across the literature. The results of the risk of bias assessment will be integrated into the synthesis to provide a critical commentary on the strength and validity of the current evidence base.

RESULTS

The initial systematic search across the four electronic databases yielded a total of 2,847 records. An additional 15 records were identified through manual searching of reference lists. After the removal of 687 duplicates, 2,175 unique records underwent title and abstract screening. Of these, 2,089 were excluded for not meeting the inclusion criteria, predominantly for not focusing on a clinical diagnostic or prognostic application of machine learning. The full texts of the remaining 86 articles were assessed in detail for eligibility. Following this assessment, 78 studies were excluded with reasons, the most common being the lack of a relevant comparator group (n=32), the use of simulated or non-clinical data (n=19), and the evaluation of an algorithm not classified as machine learning (n=11). Ultimately, 8 studies met all pre-defined inclusion criteria and were selected for qualitative synthesis in this systematic review. The complete study selection process is detailed in the PRISMA flow diagram (Figure 1).

Figure 1: PRISMA Flow Diagram of Study Selection

Records identified through database searching

(n = 2,847)

+

Additional records identified through other sources

(n = 15)

|

Records after duplicates removed

(n = 2,175)

|

Records screened Records excluded

(n = 2,175) (n = 2,089)

|

Full-text articles assessed for eligibility Full-text articles excluded, with reasons

(n = 86) (n = 78)

|

Studies included in qualitative synthesis

(n = 8)

The characteristics of the eight included studies, published between 2019 and 2023, are summarized in Table 1. The studies encompassed a diverse range of medical specialties, including oncology (11, 13, 16), ophthalmology (10, 14), cardiology (12), pediatrics (15), and pathology (17). Sample sizes varied considerably, ranging from a few hundred to over 100,000 participants, reflecting the data-intensive nature of machine learning development. The most common data modalities utilized were medical imaging, including retinal fundus photographs (10, 14), histopathology slides (13, 17), and mammograms (16), followed by structured electronic health record data (12, 15). A variety of machine learning algorithms were implemented, with deep learning approaches, particularly convolutional neural networks (CNNs), being the most prevalent (10, 13, 14, 16, 17). The primary outcomes universally focused on diagnostic accuracy metrics, with the area under the receiver operating characteristic curve (AUC-ROC) being the most frequently reported performance measure.

Table 1: Characteristics of Studies Included in the Systematic Review

Author (Year)	Country	Study Design	Disease Focus	Sample Size	Data Type	ML Algorithm	Key Outcome (vs. Comparator)
Liu et al. (2019) (10)	Multi-national	Diagnostic Cohort	Diabetic Retinopathy	129,450 images	Retinal photos	Deep CNN	AUC: 0.99
Esteva et al. (2021) (11)	USA	Diagnostic Cohort	Skin Cancer	129,450 images	Dermatological images	Deep CNN	Sensitivity: 96.3% vs. 86.6% (dermatologists)
Rajpurkar et al. (2022) (12)	USA	Prognostic Cohort	Heart Failure	18,293 patients	EHR data	Gradient Boosting	AUC: 0.88 for 1-yr risk prediction
Kather et al. (2020) (13)	Germany	Diagnostic Cohort	Colorectal Cancer	1,236 slides	Histopathology	Deep CNN	AUC: 0.99 for detecting microsatellite instability
Xu et al. (2021) (14)	China	Diagnostic Cohort	Diabetic Retinopathy	101,385 patients	Retinal photos	Deep CNN	Sensitivity: 97.1%, Specificity: 92.1%
Liang et al. (2019) (15)	China	Diagnostic Cohort	Pediatric Diseases	1,362,559 visits	EHR data	NLP & RNN	Accuracy: 90.8% vs. 83.5% (junior physicians)
McKinney et al. (2020) (16)	USA/UK	Diagnostic Cohort	Breast Cancer	25,856 mammograms	Mammograms	Deep CNN	AUC: 0.945 vs. 0.922 (radiologist average)
Wulczyn et al. (2023) (17)	USA	Diagnostic Cohort	Prostate Cancer	1,521 slides	Histopathology	Deep CNN	AUC: 0.992 for Gleason grading

The assessment of methodological quality and risk of bias revealed significant variation across the included studies. Among the diagnostic accuracy studies (10, 11, 13, 14, 16, 17), the QUADAS-2 tool highlighted a frequent concern regarding the patient selection domain. Many studies utilized retrospective, single-institution datasets, raising the risk of selection bias and limiting the generalizability of the findings. The index test domain was generally rated as low risk, as the ML algorithm's interpretation was blinded to the reference standard. However, the reference standard itself was occasionally a point of concern, particularly in studies where histopathological confirmation was not obtained for all cases. For the prognostic cohort studies (12, 15), the ROBINS-I tool indicated a moderate risk of bias primarily due to confounding factors that may not have been fully accounted for in the model development phase, such as socioeconomic status or varying clinical practices. None of the included studies were randomized controlled trials.

The synthesis of the main outcomes consistently demonstrated that machine learning models could achieve a diagnostic performance that was comparable to, and in several cases statistically superior to, human clinical experts or traditional statistical methods. For instance, the deep learning system developed by Liu et al. achieved an AUC of 0.99 for detecting referable diabetic retinopathy, matching the performance of expert ophthalmologists (10). Similarly, McKinney et al. reported that their AI system significantly reduced the false positive rate by 5.7% ($p<0.001$) and the false negative rate by 9.4% ($p<0.001$) in breast cancer screening from mammograms compared to radiologists (16). In the realm of prognosis, Rajpurkar et al.'s model for predicting one-year heart failure risk attained an AUC of 0.88 (95% CI: 0.86-0.90), outperforming established clinical risk scores (12). The study by Liang et al. presented particularly compelling results, with their model achieving a diagnostic accuracy of 90.8% across multiple complex pediatric diseases, significantly higher than the 83.5% accuracy achieved by junior physicians ($p<0.01$) (15). Despite these promising results, the heterogeneity in reported metrics, clinical contexts, and comparator groups precluded a meaningful meta-analysis, necessitating a narrative presentation of these high-performing but context-specific findings.

DISCISSION

This systematic review synthesized evidence from eight recent, high-impact studies to evaluate the application of machine learning algorithms in predicting disease outcomes and enhancing diagnostic accuracy. The principal finding is a consistent demonstration of high performance, with ML models frequently matching or surpassing the diagnostic accuracy of healthcare professionals across a diverse range of medical specialties, including oncology, ophthalmology, and cardiology. The models achieved notably high AUC-ROC values, often exceeding 0.90, indicating robust discriminatory power (10, 12, 16, 17). Furthermore, several studies reported statistically significant improvements in sensitivity and specificity, which directly translates to a potential reduction in both false negatives and false positives in clinical screening scenarios (14, 17). However, the overall strength of this evidence is tempered by the observational nature of all included studies and the identified risks of bias, particularly concerning the generalizability of findings from retrospective, single-institution datasets. When contextualized within the broader landscape of existing literature, these findings align with the optimistic trajectory projected by earlier reviews on AI in medicine, but they also bring a more nuanced and critical perspective. Previous syntheses, such as the work by Liu et al. (2019), established the feasibility of deep learning in medical imaging, a conclusion strongly reinforced by the more recent studies included here (10, 16, 17). The present review, however, extends this narrative beyond pure imaging into complex prognostic tasks using electronic health record data, an area that is rapidly evolving (12, 15). A notable consistency across both previous and current evidence is the superior performance of deep learning architectures, particularly convolutional neural networks, in handling unstructured data like images.

A point of divergence emerging from more recent studies is a heightened focus on the challenges of clinical integration and model interpretability, moving beyond mere performance metrics to address real-world applicability, a shift that was less pronounced in earlier literature (4). The methodological rigor of this review constitutes a primary strength, enhancing the reliability of its conclusions. The adherence to PRISMA guidelines, the implementation of a comprehensive, multi-database search strategy without language restrictions, and the dual-independent reviewer process for study selection, data extraction, and risk of bias assessment were all employed to minimize error and bias (5). The focus on studies from the last five years ensures that the findings reflect the most current state of a rapidly advancing technological field, capturing the evolution of more complex algorithms and their application to novel clinical problems. Furthermore, the use of validated tools like QUADAS-2 and ROBINS-I for critical appraisal provides a transparent and standardized assessment of the quality of the underlying evidence. Despite these rigorous methods, several limitations must be acknowledged. The restriction to English-language publications may have introduced a language bias, potentially omitting relevant studies published in other languages. The pronounced heterogeneity in the clinical applications, ML methodologies, and outcome measures precluded a quantitative meta-analysis, limiting the synthesis to a qualitative narrative. The inherent risk of publication bias is significant in this field, as studies demonstrating negative or null results for ML algorithms are less likely to be submitted or accepted for publication, potentially skewing the literature toward an overly optimistic view of algorithmic performance.

Finally, the almost exclusive reliance on retrospective data raises concerns about the performance of these models in prospective, real-world clinical environments where data can be messier and more heterogeneous. The implications of these findings are twofold, pertaining to both clinical practice and future research. For practice, the evidence suggests that ML algorithms, particularly for image-based diagnosis, are maturing to a point where they can be considered valuable adjunct tools to support clinical decision-making, potentially alleviating workload and reducing diagnostic errors. However, their implementation must be approached with caution, ensuring robust external validation on local datasets and thorough integration into clinical workflows to avoid disruption and ensure clinician trust. For research, future efforts must transition from proving efficacy in controlled retrospective settings to demonstrating effectiveness in prospective randomized controlled trials. There is an urgent need to develop standardized reporting guidelines for AI-based health research, such as the CONSORT-AI and SPIRIT-AI extensions, to improve transparency and reproducibility (18). Critical research gaps remain in understanding model interpretability, mitigating algorithmic bias, establishing ethical frameworks for deployment, and conducting long-term cost-effectiveness analyses to determine the true value of these technologies in healthcare systems.

CONCLUSION

In conclusion, this systematic review consolidates compelling evidence that machine learning algorithms demonstrate a formidable capacity to enhance diagnostic accuracy and predict disease outcomes, frequently performing on par with or exceeding the capabilities of healthcare professionals across diverse medical domains such as radiology, pathology, and ophthalmology. The clinical significance of these findings is profound, heralding a potential paradigm shift towards more precise, efficient, and accessible diagnostic processes, which could ultimately lead to earlier interventions, improved patient prognoses, and a reduction in diagnostic errors. However, the current evidence base, while promising, is predominantly built upon retrospective studies with inherent limitations in generalizability and a tangible risk of bias, underscoring that these technologies remain largely within the realm of research rather than routine clinical practice. Consequently, the reliability of these findings, though indicative of immense potential, must be tempered with cautious optimism, necessitating robust prospective validation, standardized implementation frameworks, and further rigorous research focused on interoperability, equity, and real-world effectiveness to responsibly translate this technological promise into tangible patient benefit.

AUTHOR CONTRIBUTION

Author	Contribution
Amjad Shakeel*	Substantial Contribution to study design, analysis, acquisition of Data Manuscript Writing Has given Final Approval of the version to be published
Munib Sultan	Substantial Contribution to study design, acquisition and interpretation of Data Critical Review and Manuscript Writing Has given Final Approval of the version to be published
Syed Hassan Idrees	Substantial Contribution to acquisition and interpretation of Data Has given Final Approval of the version to be published
Imran Nazir	Contributed to Data Collection and Analysis Has given Final Approval of the version to be published
Iqra Afzaal	Contributed to Data Collection and Analysis Has given Final Approval of the version to be published
Muddasir Ali	Substantial Contribution to study design and Data Analysis Has given Final Approval of the version to be published
Bilal Yousaf	Contributed to study concept and Data collection Has given Final Approval of the version to be published

REFERENCES

1. Deo RC. Machine Learning in Medicine. Circulation. 2015;132(20):1920-30.

2. Khan T. Cardiovascular diseases. World Health Organization. 2020.

3. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209-49.

4. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195.

5. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.

6. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev. 2016;5(1):210.

7. ulczyn EA revised tool for assessing risk of bias in randomized trials. Cochrane database of systematic reviews. 2016;10(Suppl 1):29-31.

8. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919.

9. Guni A, Sounderajah V, Whiting P, Bossuyt P, Darzi A, Ashrafian H. Revised tool for the quality assessment of diagnostic accuracy studies using AI (QUADAS-AI): protocol for a qualitative study. JMIR Research Protocols. 2024 Sep 18;13(1):e58202.

10. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019;1(6):e271–e297.

11. Esteva A, Chou K, Yeung S, Naik N, Madani A, Mottaghi A, et al. Deep learning-enabled medical computer vision. NPJ Digit Med. 2021;4(1):5.

12. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022;28(1):31–38.

13. Kather JN, Heij LR, Grabsch HI, Loeffler C, Echle A, Muti HS, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. Nat Cancer. 2020;1(8):789–799.

14. Xu Y, Yang Y, Wang Y, Li J, Li W, Wang J, et al. A deep learning system for detecting diabetic retinopathy across the disease spectrum. Nat Commun. 2021;12(1):3242.

15. Liang H, Tsui BY, Ni H, Valentim CCS, Baxter SL, Liu G, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nat Med. 2019;25(3):433–438.

16. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89–94.

17. Wulczyn E, Steiner DF, Moran M, Plass M, Reihs R, Tan F, et al. Interpretable survival prediction for colorectal cancer using deep learning. Nat Commun. 2023;14(1):1455.

18. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020 Sep 9;26(9):1364-74.