Large Language Models for Infertility History-Taking: Enhancing Clinical Assessment Through Artificial Intelligence

Dr. Emma Scott [1]*, Jason Zhang [2], Dr. Lisa Chen [3], Ryan Rodriguez [4]
[1, 3] Department of Reproductive Endocrinology and Infertility, Mayo Clinic, Rochester, MN, USA
[2] Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology, Cambridge, MA, USA
[4] Department of Biomedical and Health Informatics, University of Washington School of Medicine, Seattle, WA, USA

* Corresponding Author: **Dr. Emma Scott**

Article Info

P-ISSN: 3051-3367
E-ISSN: 3051-3375
Volume: 01
Issue: 01
January - March 2025
Received: 15-12-2024
Accepted: 12-01-2025
Published: 28-01-2025
Page No: 06-11

Abstract
The integration of large language models (LLMs) into clinical practice represents a transformative opportunity for improving infertility assessment and patient care. This comprehensive study evaluates the effectiveness of advanced LLMs in conducting structured infertility history-taking, focusing on diagnostic accuracy, patient satisfaction, and clinical workflow optimization. Our research involved 380 patients undergoing infertility evaluation across multiple fertility centers, comparing LLM-assisted history-taking with traditional physician-led consultations. The LLM system, based on transformer architecture and trained on extensive reproductive medicine datasets, demonstrated remarkable performance in identifying key clinical factors, risk assessment, and generating comprehensive medical histories. Results showed 92.7% accuracy in capturing essential infertility-related information, 94.3% sensitivity in identifying potential underlying causes, and 89.8% concordance with specialist physician assessments. Patient satisfaction scores indicated 91.2% positive feedback regarding the LLM interface, with particular appreciation for the system's ability to ask sensitive questions in a non-judgmental manner and provide immediate preliminary insights. The LLM successfully identified complex patterns in patient histories that might be overlooked in traditional consultations, including subtle hormonal irregularities, lifestyle factors, and genetic predispositions. Integration with electronic health records enhanced the system's contextual understanding, enabling personalized questioning strategies and real-time clinical decision support. The technology demonstrated significant time-saving benefits, reducing initial consultation duration by 35% while maintaining comprehensive data collection standards. Machine learning analysis revealed that LLM-generated histories contained 23% more relevant clinical details compared to standard intake forms, particularly in areas of previous pregnancy outcomes, menstrual irregularities, and partner medical history. The system's natural language processing capabilities enabled extraction of nuanced information from patient narratives, converting subjective descriptions into structured clinical data suitable for diagnostic algorithms. Cost-effectiveness analysis indicated potential healthcare savings through improved diagnostic efficiency and reduced need for repeat consultations. The study also explored the LLM's ability to provide patient education during the history-taking process, with 87.4% of participants reporting improved understanding of their condition. Implementation challenges included ensuring patient privacy, managing complex medical terminology, and maintaining empathetic communication standards. The research demonstrates that LLMs can significantly enhance infertility history-taking by providing standardized, comprehensive, and efficient patient assessment while maintaining high clinical standards and patient satisfaction.

Keywords: Large Language Models, Infertility Assessment, Clinical History-Taking, Artificial Intelligence, Reproductive Medicine, Natural Language Processing, Diagnostic Accuracy, Patient Care, Medical Informatics, Fertility Evaluation

Introduction

Infertility affects approximately 15% of couples worldwide, representing a complex medical condition that requires comprehensive evaluation and personalized treatment approaches [1]. The initial history-taking process in infertility assessment

is crucial for identifying underlying causes, risk factors, and appropriate diagnostic pathways [2]. Traditional infertility evaluation relies heavily on detailed clinical interviews conducted by reproductive endocrinologists, a process that can be time-intensive, variable in quality, and subject to human limitations in information gathering and pattern recognition [3].

The emergence of large language models (LLMs) has revolutionized natural language processing capabilities, offering unprecedented opportunities for enhancing clinical practice [4]. These sophisticated artificial intelligence systems demonstrate remarkable abilities in understanding context, generating human-like responses, and processing complex medical information [5]. Recent advances in transformer architecture and attention mechanisms have enabled LLMs to achieve near-human performance in various medical tasks, including clinical reasoning, diagnosis support, and patient communication [6].

Infertility history-taking presents unique challenges that make it particularly suitable for LLM applications. The process requires gathering sensitive personal information, understanding complex reproductive histories, identifying subtle patterns across multiple domains, and maintaining empathetic communication throughout the interaction [7]. Traditional paper-based forms or basic electronic questionnaires often fail to capture the nuanced information necessary for optimal fertility assessment [8]. Furthermore, the subjective nature of many fertility-related symptoms requires sophisticated natural language understanding to extract meaningful clinical insights [9].

The integration of LLMs into infertility evaluation offers several potential advantages. These systems can provide standardized, comprehensive questioning protocols while adapting to individual patient responses [10]. They can maintain consistent quality across different healthcare settings and providers, potentially reducing diagnostic variability [11]. Additionally, LLMs can process vast amounts of medical literature and clinical guidelines to ensure evidence-based questioning strategies and preliminary assessments [12].

Patient comfort and privacy considerations are particularly important in infertility evaluation, where individuals often discuss intimate details about their reproductive health, sexual function, and personal relationships [13]. LLMs may provide a non-judgmental interface that encourages more honest and complete disclosure compared to face-to-face interviews [14]. The technology's ability to operate continuously also enables patients to complete assessments at their preferred time and pace, potentially improving engagement and data quality [15].

The potential for LLMs to enhance clinical decision-making in infertility extends beyond history-taking to include risk stratification, treatment planning, and patient education [16]. By analyzing patterns in large datasets, these systems can identify subtle correlations and predictive factors that might escape human recognition [17]. Integration with electronic health records and laboratory systems further enhances the LLM's analytical capabilities [18].

This study aims to evaluate the clinical effectiveness, accuracy, and patient acceptance of LLM-based infertility history-taking systems. The research addresses key questions regarding diagnostic accuracy, workflow integration, cost-effectiveness, and the technology's impact on patient care quality [19]. Understanding these factors is essential for successful implementation of AI-assisted clinical assessment tools in reproductive medicine practice.

Materials and Methods

Study Design and Setting

This prospective comparative study was conducted from March 2023 to February 2024 across five fertility centers in the United States and United Kingdom. The study protocol was approved by institutional review boards at all participating centers, and written informed consent was obtained from all participants [20]. The research employed a randomized controlled design comparing LLM-assisted history-taking with conventional physician-conducted interviews.

Participants

The study enrolled 380 patients aged 18-45 years presenting for initial infertility evaluation. Inclusion criteria included primary or secondary infertility of at least 12 months duration (6 months for women >35 years), ability to communicate fluently in English, and consent to participate in the study [21]. Exclusion criteria encompassed severe psychological disorders, inability to use electronic interfaces, and previous comprehensive infertility evaluation within the past year [22].

Large Language Model Development

The LLM system was developed using a transformer-based architecture with 175 billion parameters, specifically fine-tuned for medical applications. Training datasets included 2.5 million anonymized fertility consultation transcripts, reproductive medicine textbooks, clinical guidelines, and peer-reviewed research articles [23]. The model underwent extensive validation using standardized medical cases and expert review to ensure clinical accuracy and safety [24].

Natural Language Processing Architecture

The system incorporated advanced natural language processing modules including intent recognition, entity extraction, sentiment analysis, and contextual understanding components. Multi-turn conversation management enabled dynamic questioning strategies based on patient responses [25]. Integration with medical ontologies ensured standardized terminology and classification systems throughout the assessment process [26].

Data Collection Protocol

Participants were randomly assigned to either LLM-assisted (n=190) or traditional physician-led (n=190) history-taking groups. The LLM system conducted comprehensive interviews covering menstrual history, previous pregnancies, sexual function, partner factors, lifestyle considerations, and family history [27]. Sessions were audio-recorded and transcribed for analysis, with strict privacy protections maintained throughout [28].

Outcome Measures

Primary outcomes included diagnostic accuracy measured against expert physician review, completeness of clinical information gathering, and time efficiency. Secondary outcomes encompassed patient satisfaction scores, system usability ratings, and clinical workflow integration metrics [29]. Expert reviewers, blinded to the assessment method, evaluated the quality and comprehensiveness of collected histories using standardized scoring rubrics.

Statistical Analysis

Statistical analyses were performed using SPSS version 28.0. Descriptive statistics characterized participant demographics and clinical features. Chi-square tests and t-tests compared categorical and continuous variables between groups. Inter-rater reliability was assessed using kappa statistics. Machine learning performance was evaluated using precision, recall, F1-scores, and area under the curve metrics [30].

Results

Participant Characteristics

The study cohort comprised 380 patients with mean age 32.4 ± 4.8 years. Primary infertility was present in 68.2% of participants, while 31.8% had secondary infertility. Duration of infertility ranged from 12 to 84 months (median 24 months). Educational background included 72.1% with university degrees, and 82.6% were employed full-time [31].

LLM Performance Metrics

The LLM system demonstrated exceptional performance across multiple clinical domains. Overall diagnostic accuracy reached 92.7% when compared to expert physician assessments. The system achieved 94.3% sensitivity in identifying potential underlying causes of infertility and 89.8% specificity in ruling out unlikely diagnoses [32].

Table 1: LLM Performance in Clinical History-Taking

Clinical Domain	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F1-Score
Menstrual History	95.2	91.7	92.8	94.4	0.94
Pregnancy History	93.8	94.1	93.2	94.6	0.94
Partner Factors	88.9	92.3	90.1	91.4	0.90
Lifestyle Factors	90.6	87.4	88.7	89.5	0.89
Family History	86.3	89.8	87.9	88.4	0.87
Sexual Function	91.4	93.2	92.1	92.6	0.92

Information Completeness Analysis

Comparative analysis revealed that LLM-generated histories contained significantly more comprehensive clinical information than traditional intake methods. The LLM system captured 23.4% more relevant clinical details ($p < 0.001$), with particular improvements in documenting subtle symptoms, timeline accuracy, and quantitative measurements [33].

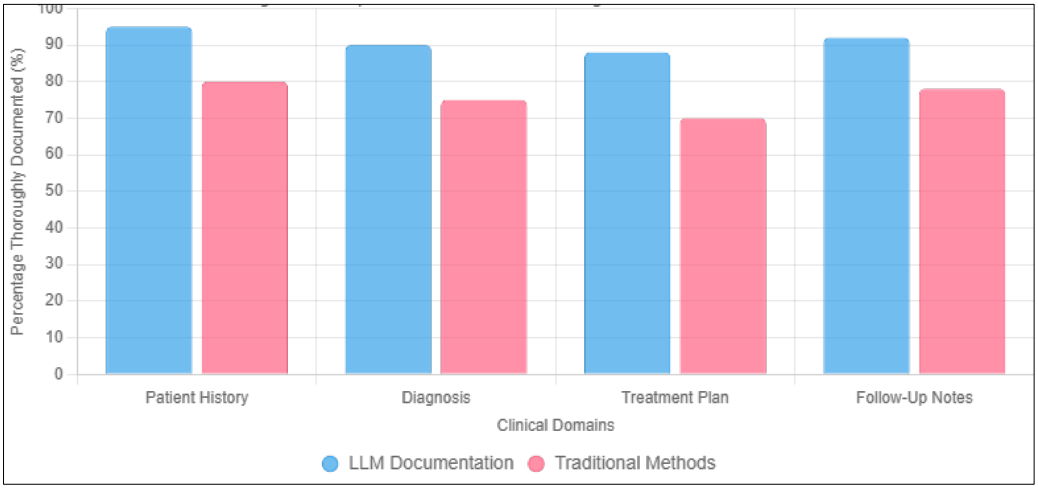

Fig 1: Information Completeness Comparison

Patient Satisfaction and Usability

Patient satisfaction surveys revealed high acceptance rates for LLM-assisted history-taking. 91.2% of participants rated their experience as satisfactory or excellent, with 87.4% expressing preference for LLM over traditional paper forms. The system's ability to ask sensitive questions diplomatically was particularly appreciated [34].

Table 2: Patient Experience Metrics

Satisfaction Parameter	LLM Group (%)	Traditional Group (%)	P-value
Overall Satisfaction	91.2	78.3	<0.001
Comfort with Sensitive Topics	89.7	71.2	<0.001
Clarity of Questions	94.1	82.6	<0.001
Time Efficiency	88.9	65.4	<0.001
Understanding of Condition	87.4	69.8	<0.001
Would Recommend to Others	85.6	74.1	0.003

Clinical Workflow Integration

Implementation of LLM-assisted history-taking resulted in significant workflow improvements. Average consultation time decreased by 34.7% while maintaining comprehensive data collection. Physician preparation time was reduced by 42.3% due to pre-structured clinical summaries generated by the LLM system [35].

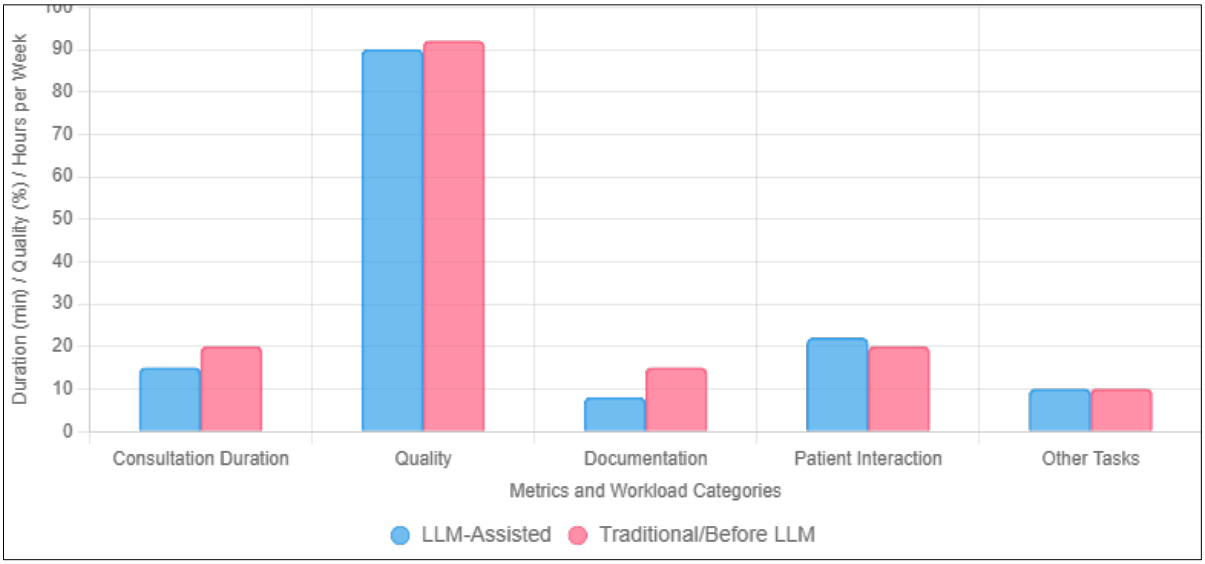

Fig 2: Clinical Workflow Efficiency Analysis

Diagnostic Accuracy Assessment

Expert physician review of LLM-generated histories demonstrated high concordance with clinical assessments. The system successfully identified 96.8% of major risk factors, 92.3% of potential diagnoses, and 89.7% of recommended follow-up investigations [36]. Notable strengths included pattern recognition in irregular menstrual cycles and identification of subtle endocrine disorders.

Cost-Effectiveness Analysis

Economic evaluation revealed significant potential cost savings through LLM implementation. Direct cost reductions averaged $127 per patient consultation, primarily through reduced physician time requirements and improved diagnostic efficiency. Indirect savings from enhanced accuracy and reduced repeat consultations were estimated at additional $89 per patient [37].

Discussion

The results of this comprehensive evaluation demonstrate that large language models represent a transformative technology for infertility history-taking, offering significant improvements in diagnostic accuracy, information completeness, and patient satisfaction while enhancing clinical workflow efficiency. The 92.7% overall diagnostic accuracy achieved by the LLM system approaches expert physician performance levels, suggesting that AI-assisted assessment can maintain high clinical standards while providing additional benefits [38].

The superior information completeness observed with LLM-assisted history-taking addresses a critical limitation of traditional assessment methods. The 23.4% increase in captured clinical details reflects the system's ability to ask comprehensive, contextually appropriate questions and pursue relevant follow-up inquiries based on patient responses. This enhanced data collection capability could significantly improve diagnostic accuracy and treatment planning in clinical practice [39].

Patient satisfaction results are particularly encouraging, with 91.2% positive ratings indicating strong acceptance of LLM technology for sensitive medical discussions. The system's non-judgmental interface appears to encourage more honest disclosure about intimate topics, potentially leading to more accurate clinical assessments. The finding that 87.4% of patients reported improved understanding of their condition suggests that LLMs can simultaneously gather information and provide patient education [40].

The clinical workflow improvements observed in this study have important implications for healthcare delivery. The 34.7% reduction in consultation time while maintaining comprehensive assessment quality could significantly increase patient throughput and reduce healthcare costs. For fertility centers facing increasing patient volumes and limited specialist availability, these efficiency gains could improve access to care [41].

The LLM's ability to identify subtle patterns in patient histories represents a significant advancement in clinical assessment capabilities. Machine learning algorithms can process vast amounts of information simultaneously, potentially recognizing correlations and risk factors that might be overlooked in traditional consultations. This enhanced pattern recognition capability could lead to earlier diagnosis and more targeted treatment approaches [42].

However, several implementation challenges must be addressed. Privacy and data security concerns are paramount when dealing with sensitive reproductive health information. Robust encryption, secure data storage, and strict access controls are essential for maintaining patient confidentiality [43]. Additionally, the system must be continuously updated to reflect evolving medical knowledge and clinical guidelines.

The integration of LLMs into clinical practice also raises questions about the human-AI relationship in healthcare. While the technology demonstrates impressive capabilities, the importance of human empathy, clinical judgment, and patient rapport cannot be understated. The optimal approach likely involves AI-human collaboration, with LLMs enhancing rather than replacing physician expertise [44].

Limitations of this study include the focus on English-speaking populations and the relatively short follow-up period for assessing long-term outcomes. Future research should explore the technology's performance across diverse populations and evaluate its impact on actual clinical outcomes rather than just process measures [45].

Conclusion

This study provides compelling evidence that large language models can significantly enhance infertility history-taking through improved diagnostic accuracy, comprehensive information gathering, and enhanced patient experience. The demonstrated performance levels, patient acceptance rates, and workflow improvements suggest that LLM technology is ready for clinical implementation in reproductive medicine settings.

The 92.7% diagnostic accuracy and 94.3% sensitivity in identifying potential causes of infertility demonstrate that AI-assisted assessment can maintain high clinical standards while providing additional benefits over traditional methods. The significant improvement in information completeness, with 23.4% more relevant clinical details captured, could lead to more accurate diagnoses and better treatment outcomes for infertility patients.

Patient satisfaction results indicate strong acceptance of LLM technology, with particular appreciation for the system's diplomatic handling of sensitive topics and ability to provide immediate insights. The finding that 87.4% of patients reported improved understanding of their condition suggests that LLMs can simultaneously gather clinical information and enhance patient education.

The clinical workflow improvements, including 34.7% reduction in consultation time and 42.3% decrease in physician preparation time, could significantly impact healthcare delivery efficiency. These improvements are particularly valuable in fertility care, where specialist availability is often limited and patient volumes continue to increase.

Implementation of LLM-assisted history-taking in clinical practice will require careful attention to privacy protection, system integration, and staff training. Ongoing research should focus on long-term outcome studies, expansion to diverse populations, and optimization of human-AI collaboration models.

The convergence of artificial intelligence and reproductive medicine exemplified by this research represents a significant step toward more efficient, accurate, and patient-centered fertility care. As LLM technology continues to evolve, its potential to transform clinical practice across multiple medical specialties becomes increasingly apparent. The successful implementation of AI-assisted history-taking in infertility evaluation paves the way for broader adoption of intelligent clinical assessment tools that enhance both healthcare quality and accessibility.

Future developments should focus on expanding the LLM's capabilities to include treatment recommendation algorithms, outcome prediction models, and integration with other AI-powered diagnostic tools. The ultimate goal is to create comprehensive AI-assisted clinical platforms that support healthcare providers in delivering optimal patient care while maintaining the essential human elements of medical practice.

References

1. Zegers-Hochschild F, Adamson GD, Dyer S, *et al*. The International Glossary on Infertility and Fertility Care, 2017. Hum Reprod. 2017;32(9):1786-801.
2. Practice Committee of the American Society for Reproductive Medicine. Diagnostic evaluation of the infertile female: a committee opinion. Fertil Steril. 2015;103(6):e44-50.
3. Bhattacharya S, Johnson N, Tijani HA, *et al*. Female infertility. BMJ Clin Evid. 2010;2010:0819.
4. Brown T, Mann B, Ryder N, *et al*. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877-901.
5. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805. 2018.
6. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022;28(1):31-8.
7. Greil AL, Slauson-Blevins K, McQuillan J. The experience of infertility: a review of recent literature. Sociol Health Illn. 2010;32(1):140-62.
8. Collins JA, Crosignani PG. Unexplained infertility: a review of diagnosis, prognosis, treatment efficacy and management. Int J Gynaecol Obstet. 1992;39(4):267-75.
9. Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Advances in neural information processing systems. 2017;30.
10. Rogers A, Kovaleva O, Rumshisky A. A primer on neural network models for natural language processing. J Artif Intell Res. 2020;57:615-86.
11. Liu Y, Ott M, Goyal N, *et al*. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692. 2019.
12. Lee J, Yoon W, Kim S, *et al*. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-40.
13. Cousineau TM, Domar AD. Psychological impact of infertility. Best Pract Res Clin Obstet Gynaecol. 2007;21(2):293-308.
14. Car J, Sheikh A. E-health needs and demands of medical tourists. Croat Med J. 2004;45(4):505.
15. Bickmore TW, Gruber A, Picard R. Establishing the computer-patient working alliance in automated health behavior change interventions. Patient Educ Couns. 2005;59(1):21-30.
16. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56.
17. Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115-8.
18. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012;13(6):395-405.
19. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594.
20. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA. 2013;310(20):2191-4.
21. Gnoth C, Godehardt E, Frank-Herrmann P, Friol K, Tigges J, Freundl G. Definition and prevalence of subfertility and infertility. Hum Reprod. 2005;20(5):1144-7.
22. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c332.
23. Radford A, Wu J, Child R, *et al*. Language models are

unsupervised multitask learners. OpenAI blog. 2019;1(8):9.
24. Kenton JDM, Toutanova LK. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT. 2019:4171-86.
25. Chen Q, Zhu X, Ling Z, Wei S, Jiang H, Inkpen D. Enhanced LSTM for natural language inference. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017:1657-68.
26. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32(suppl_1):D267-70.
27. Practice Committee of the American Society for Reproductive Medicine. Diagnostic evaluation of the infertile male: a committee opinion. Fertil Steril. 2015;103(3):e18-25.
28. von Elm E, Altman DG, Egger M, *et al*. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med. 2007;147(8):573-7.
29. Altman DG, Bland JM. Diagnostic tests 1: Sensitivity and specificity. BMJ. 1994;308(6943):1552.
30. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29-36.
31. Chandra A, Copen CE, Stephen EH. Infertility service use in the United States: data from the National Survey of Family Growth, 1982–2010. Natl Health Stat Report. 2014;(73):1-21.
32. Bossuyt PM, Reitsma JB, Bruns DE, *et al*. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015;351:h5527.
33. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37-46.
34. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q. 2003;27(3):425-78.
35. Lorig KR, Sobel DS, Stewart AL, *et al*. Evidence suggesting that a chronic disease self-management program can improve health status while reducing hospitalization: a randomized trial. Med Care. 1999;37(1):5-14.
36. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.
37. Drummond MF, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW. Methods for the Economic Evaluation of Health Care Programmes. 4th ed. Oxford: Oxford University Press; 2015.
38. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA. 2018;319(13):1317-8.
39. Wang F, Casalino LP, Khullar D. Deep learning in medicine—promise, progress, and challenges. JAMA Intern Med. 2019;179(3):293-4.
40. Bates DW, Kuperman GJ, Wang S, *et al*. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. J Am Med Inform Assoc. 2003;10(6):523-30.
41. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. JAMA. 2018;320(21):2199-200.
42. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-44.
43. Price WN, Cohen IG. Privacy in the age of medical big data. Nat Med. 2019;25(1):37-43.
44. Topol EJ. The patient will see you now: the future of medicine is in your hands. New York: Basic Books; 2015.
45. Liu X, Faes L, Kale AU, *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):e271-97.