International Journal of Research in Medical Science
www.medicalpaper.net
Online ISSN: 2664-8741, Print ISSN: 2664-8733
Received: 15-11-2021, Accepted: 05-12-2021, Published: 16-12-2021
Volume 3, Issue 2, 2021, Page No. 14-19

AI in mental health diagnostics: Ethical imperatives and design strategies for equitable implementation

Feyikemi Mary Akinyelure
Federal School of Occupational Therapy, Oshodi, Nigeria

DOI: https://www.doi.org/10.33545/26648733.2021.v3.i2a.167

Abstract
Artificial Intelligence (AI) is increasingly being used in mental health diagnostics, creating new opportunities for early identification, higher accuracy, and more access to care. Natural language processing (NLP), deep learning models, and chatbot-based evaluations are being used to analyze voice, text, facial expressions, and behavioural patterns in order to detect disorders such as depression, anxiety, and schizophrenia. Despite these advances, the use of AI in mental health presents serious ethical and equitable concerns, specifically over data bias, transparency, cultural sensitivity, and the potential of exacerbating gaps in mental health care. This review aims to explore the ethical dimensions and equity implications of AI-based tools in mental health diagnostics. The findings indicate that, while AI has diagnostic accuracy in controlled conditions, its generalizability to different populations is limited due to non-representative training datasets and a lack of cultural and contextual adaptation. Ethical issues in AI integration, including algorithmic opacity, data privacy, and the digital divide, hinder real-world integration. Strategies such as explainable AI, fairness-aware modelling, and participatory design offer potential solutions. Therefore, addressing ethical issues and structural constraints through collaborative, transparent, and context-sensitive design is critical to preventing existing gaps.

Keywords: AI-driven approaches, digital mental health, equitable implementation, health equity, data privacy

1. Introduction

The global mental health crisis is a growing problem, with diseases such as depression, anxiety, bipolar disorder, and schizophrenia impacting people globally (Colizzi *et al*., 2020) [8]. Mental health disorders are increasingly a primary source of disability, contributing significantly to the global burden of disease. Despite their widespread prevalence, diagnosing mental health conditions remains challenging. Conventional diagnostic methods rely heavily on subjective clinical judgement, patient self-reporting, and standardized assessment tools, which may be insensitive to cultural and individual differences (Clark *et al*., 2017) [7]. Therefore, considerable discrepancies in access to mental healthcare persist, particularly in low-income and marginalized communities.

Artificial intelligence (AI) has emerged as a tool capable of transforming and reshaping the field of mental health diagnosis. AI-powered technologies such as machine learning algorithms, natural language processing (NLP), and predictive analytics provide unprecedented capabilities for detecting, classifying, and monitoring psychological conditions based on speech, text, facial expressions, biometric sensors, and digital behaviours (McKinsey & Company, 2020) [24]. These technologies can potentially improve diagnostic precision, allow for early detection, and increase access to care via automated or remote systems. However, the use of AI in mental health presents major ethical and equitable concerns (Janssen *et al*., 2018) [18]. Therefore, issues such as algorithmic bias, a lack of transparency, data privacy violations, and unequal access to digital health technologies have the potential to perpetuate or exacerbate existing gaps in mental health diagnosis and treatment.

AI technologies have shown progress in detecting mental health disorders; however, many of the systems are trained on non-representative datasets, lack cultural and contextual sensitivity, and are developed without meaningful involvement with end users or vulnerable communities (Graham *et al*., 2019) [14]. In this regard, AI systems may misdiagnose individuals, neglect unique symptom presentations, or marginalize those not conforming to the normative data profiles. This is particularly problematic in mental health, where stigma, structural barriers, and social determinants have a vital impact on diagnosis and treatment access (Stangl *et al*., 2019) [30]. These limitations highlight the urgent need for scalable, objective, and inclusive diagnostic technologies that can enhance clinical capacity and improve the accuracy of mental health assessments. Therefore, this review explores the ethical dimensions and equity implications of AI-based tools in mental health diagnostics.

2. AI in Mental Health Diagnostics

AI is emerging as a force in mental health diagnostics, promising more objective, scalable, and rapid diagnosis of cognitive diseases. Unlike many physical health conditions, which can be diagnosed using measurable biomarkers like blood tests or imaging, mental health diagnoses are currently based on subjective assessments,

clinical interviews, and self-reported symptoms (Ternes *et al*., 2020) [31]. These methods, while useful, are prone to bias, inconsistencies, and diagnostic delays, especially when physicians are overwhelmed or patients lack the linguistic or cultural competence to communicate their suffering in ways that correspond to standardized diagnostic criteria (Soled, 2020) [28]. In contrast, AI allows for the systematic examination of massive and diverse data sources to identify patterns related to mental health disorders that may be difficult for humans to detect (Ogundare, 2019) [26]. This capability establishes AI as a powerful tool for enhancing clinical judgement, reducing diagnostic variability, and increasing access to mental health assessments, particularly in underserved communities.

2.1 Emerging Evidence for Diagnostic Accuracy

Research into AI-based mental health diagnoses has produced findings, particularly in the areas of depression, anxiety, schizophrenia, and related illnesses. Several studies have found that AI models can attain diagnostic accuracies comparable to or even higher than those of qualified healthcare professionals under controlled conditions (Kuziemsky *et al*., 2019; Carr, 2020) [20, 4]. Specifically, Natural Language Processing (NLP) algorithms analyzing patient narratives or clinical transcripts have been shown to detect depressive symptoms with an accuracy by identifying patterns such as reduced lexical diversity, negative sentiment, and changes in syntactic structure (Alanazi *et al*., 2017) [1]. Similarly, voice-based AI models trained to assess tone, pitch, and rhythm have demonstrated high sensitivity and specificity in distinguishing manic from depressive episodes in bipolar disorder. In schizophrenia, machine learning algorithms that analyze disorganized speech and semantic coherence have shown strong predictive performance in early detection studies, identifying individuals at risk for psychosis months before clinical symptoms appear (Guidi *et al*., 2017) [15].

Furthermore, multimodal AI systems, which incorporate text, voice, visual signals, and behavioural data (such as sleep, movement, and smartphone use), have demonstrated even higher diagnostic potential. These systems are capable of recognizing complicated mental health conditions that single-modality techniques may ignore (Garcia-Ceja *et al*., 2018) [12]. AI-powered chatbots incorporated in mobile mental health apps have been tested in randomized controlled trials, with some demonstrating considerable gains in the early detection of anxiety and depression symptoms, particularly among young adults and adolescents (Fulmer *et al*., 2018) [11]. This, therefore, implies that, given the right conditions, AI tools can play an important role in increasing access to mental health diagnostics while also enhancing assessment accuracy and speed.

Table 1: Summary of AI-based diagnostic modalities and their reported accuracy in mental health disorders

Mental Health Disorder	AI Modality / Technique	Primary Data Source	Reported Diagnostic Accuracy	Key Findings / Remarks	Reference
Depression	Natural Language Processing (NLP)	Clinical transcripts, patient narratives	85-92%	NLP models detect depressive symptoms via reduced lexical diversity, negative sentiment, and syntactic changes.	Alanazi *et al*., 2017 [1]
Anxiety Disorders	AI-based Chatbots	Text-based mobile applications	78-88%	Chatbots demonstrate early detection and support among young adults; improved screening speed and accessibility.	Fulmer *et al*., 2018 [11]
Bipolar Disorder	Voice and Speech Analysis	Audio recordings (tone, pitch, rhythm)	80-90%	Voice-based AI distinguishes manic from depressive episodes with high sensitivity and specificity.	Carr, 2020 [4]
Schizophrenia	Machine Learning Speech Models	Clinical interviews and speech samples	83-91%	Algorithms detect disorganized speech and semantic incoherence for early psychosis prediction.	Guidi *et al*., 2017 [15]
Multimodal Systems	Integrated Text, Voice, and Behavioural Analytics	Smartphone sensors, wearables, and EHR data	90-95%	Combining behavioural, linguistic, and physiological data enhances overall diagnostic precision.	Garcia-Ceja *et al*., 2018 [12]
General Mental Health Screening	Ensemble Deep Learning Models	Mixed clinical and self-reported datasets	87-93%	AI tools perform comparably or exceed clinician accuracy under controlled conditions.	Kuziemsky *et al*., 2019 [20]

The table provides an overview of key artificial intelligence applications across major mental health conditions, including depression, anxiety, bipolar disorder, and schizophrenia. It summarizes the diagnostic techniques such as Natural Language Processing (NLP), voice and speech analysis, and multimodal systems alongside their reported accuracies, data sources, and clinical validation outcomes. The table highlights comparative performance between single-modality and multimodal AI systems, illustrating how text, voice, and behavioural data integration enhances predictive sensitivity and specificity in early detection, as demonstrated in studies by Kuziemsky, Carr, Alanazi, Guidi, Garcia-Ceja, and Fulmer.

2.2 Challenges in Explainability and Trust

Another restriction is the lack of transparency in AI decision-making processes, particularly in deep learning models. Many AI systems function as black boxes, producing diagnostic results without providing explicit explanations for how conclusions are made. In mental health care, where trust, empathy, and clinical reasoning are essential components of the therapeutic interaction, a lack of explainability can undermine physician acceptance and patient confidence (Fitzpatrick *et al*., 2017) [10]. Clinicians are often hesitant to rely on tools that cannot be interrogated or contextualized, particularly when making sensitive or potentially life-altering decisions such as starting psychiatric medication or admitting a patient to the hospital (Chin-Yee & Upshur, 2020; Dousa, 2020) [6, 9]. Even the most accurate AI systems may go underutilized in clinical settings if there are no processes in place to ensure transparent and interpretable output.

2.3 Ethical and practical constraints

Ethical concerns about data privacy, consent, and algorithmic accountability further limit the use of AI in mental health diagnostics. Many systems rely on passive collection of behavioural and biometric data via smartphones, wearable's, or digital platforms, raising issues about surveillance and autonomy, especially when users are unaware of the data being collected or cannot provide informed consent (Maher *et al*., 2019) [22]. Furthermore, the legal and regulatory frameworks for AI in healthcare remain underdeveloped, especially in low- and middle-income countries. This causes uncertainties for providers and developers around liability, quality assurance, and compliance (Gerke *et al*., 2020) [13]. Therefore, without strong governance structures and clear guidelines for AI implementation, health systems may struggle to adopt these technologies safely and fairly.

3. Health Equity and the Risk of Exacerbating Disparities

The incorporation of AI into mental health diagnostics holds the possibility of increased access to care, enhanced diagnostic precision, and real-time monitoring, especially in resource-constrained settings. However, this promise is tempered by serious concerns about health equity and the possibility of exacerbating existing disparities in mental health care (Chen *et al*., 2019) [5]. These problems originate from the fact that AI systems are only as good as the data on which they are trained, and much of the available training data reflects historical biases, structural inequalities, and under-representation of marginalized people. The possibility of algorithmic bias is especially concerning in the field of mental health, since diagnostic criteria are frequently subjective and culturally influenced. For example, emotional expressiveness, help-seeking behaviour, and symptom disclosure differ greatly between cultures and societies (Ngiam & Khor, 2019) [25]. Similarly, language models that evaluate depressive or psychotic speech patterns may interpret dialectal or non-standard language usage as incoherence or disorganization. These cultural mismatches between model training data and real-world user expression might result in false positives, over-pathologization, or failure to identify individuals in true need of care, perpetuating rather than reducing diagnostic disparities (Komeili *et al*., 2019) [19].

Furthermore, employers or universities may use AI tools to screen for psychological fitness or behavioural risks without providing adequate transparency, consent, or follow-up care. This not only violates ethical standards, but it may disproportionately target or stigmatize people from marginalized communities, reinforcing existing monitoring and discrimination systems. In such cases, AI becomes a vehicle of exclusion and injury rather than a tool for empowerment or equity (Malik *et al*., 2019) [23]. To avoid these results, AI for mental health diagnostics must be designed and used with conscious, equity-focused tactics. This involves diversifying training datasets to account for a wide range of cultural, linguistic, and demographic realities; incorporating under-represented communities in AI tool design and evaluation; and undertaking fairness audits to detect and correct algorithmic prejudice (Padrez *et al*., 2015) [27]. Developers must also prioritize openness, explainability, and accountability, ensuring that users understand how diagnostic decisions are made and that means exist to debate or correct errors. Governments and regulatory agencies play a critical role in creating ethical standards and supervision systems that ensure equal access and protect against misuse (Jameel *et al*., 2020) [17]. Therefore, while AI has the potential to democratize access to mental health care, it will only be realized if equity is included in every stage of its development and deployment.

4. Strategies for Equitable and Ethical AI Systems

In order to ensure that AI systems in mental health diagnostics are effective, equitable, and ethically sound, they must be developed using a comprehensive set of design techniques. These tactics must enable algorithmic performance and consider the human, cultural, and systemic settings in which these technologies will be used (Stahl & Wright, 2018) [29]. Human-centered design, participatory innovation, bias mitigation, transparency, and interdisciplinary collaboration are critical for developing systems that promote mental health equity, protect vulnerable populations, and adhere to healthcare ethics.

4.1 Human-centered and Participatory Design

Human-centered and participative design is at the heart of equitable AI development, putting end-users' requirements, experiences, and values first, particularly those from marginalized populations. This strategy shifts away from top-down, technocratic innovation approaches and towards iterative engagement with stakeholders such as patients, carers, mental health professionals, and community leaders (Auernhammer, 2020) [3]. By actively incorporating these groups in the design process via interviews, co-design workshops, and user testing, developers can find context-specific needs, usability impediments, and culturally relevant mental health indicators.

4.2 Bias Mitigation Techniques for AI Models

AI models are susceptible to biases resulting from historical inequities, data imbalances, and algorithmic design decisions. These biases can result in systematic misdiagnosis, over diagnosis, or the exclusion of specific populations from diagnostic pathways. Therefore, to solve this issue, developers must integrate bias mitigation measures throughout the AI development process (Lee *et al*., 2019) [21].

4.3 Transparent and Explainable AI (XAI)

In clinical and diagnostic settings, trust and accountability are critical. AI systems must not only generate accurate results, but also provide clear, understandable reasoning for their outputs, particularly when used to assist sensitive judgements such as mental health diagnosis (Amann *et al*., 2020) [2]. Explainable AI (XAI) is especially useful in this situation. XAI tools provide insights into how AI models arrive at certain conclusions, either by highlighting which features were most relevant in a given prediction or by breaking down complex models into digestible

components. Attention mechanisms in neural networks enable physicians and patients to question AI judgements, identify flaws, and contextualize outcomes within clinical judgement (Holzinger *et al*., 2019) [16].

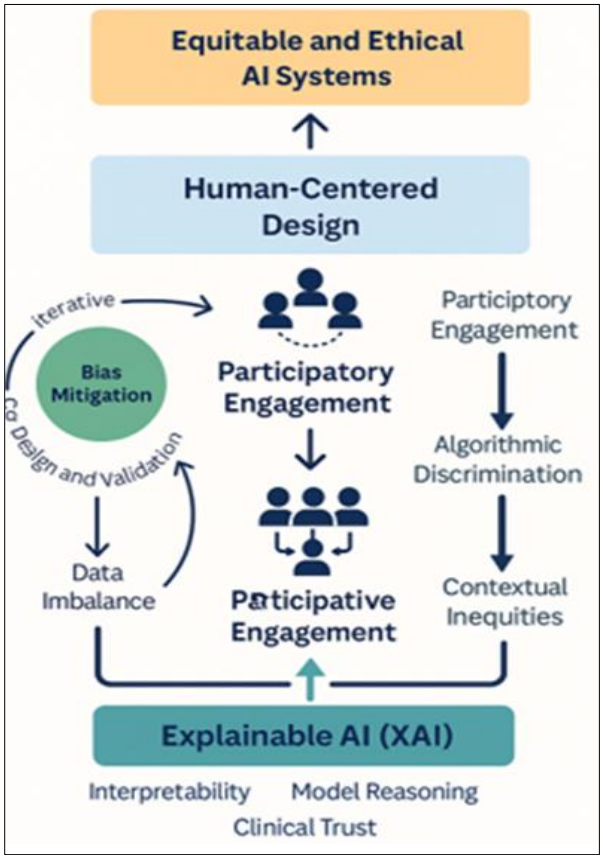

Fig 1: Framework for developing equitable and ethical AI systems in mental health diagnostics.

5. Conclusion

The introduction of AI in mental health diagnostics is a breakthrough point in the evolution of psychiatric treatment, with the potential to transform how mental disorders are recognized, monitored, and controlled. AI provides unprecedented opportunities to improve diagnostic accuracy, streamline clinical workflows, and increase access to care, particularly in regions or populations that have historically been underserved by conventional mental health services. Natural language processing tools detect subtle linguistic indicators of depression, while facial recognition algorithms observe affective expressions. However, from this review, these technological improvements are followed by ethical, social, and practical issues that must be addressed by thoughtful and purposeful design.

The notion of equity, inclusion, and openness is crucial to the proper deployment of AI in mental health diagnosis. AI systems trained on non-representative data or deployed without cultural sensitivity risk perpetuating or exacerbating long-standing gaps in mental health access and outcomes. Marginalized groups, such as racial minorities, non-English speakers, rural communities, and those with poor digital access, are most vulnerable to exclusion or harm if AI systems fail to reflect their experiences. As a result, guaranteeing equal outcomes necessitates incorporating fairness into all stages of the AI lifecycle, from dataset creation and model training to deployment, evaluation, and post-deployment monitoring.

The review identified essential techniques for guiding the ethical development and implementation of AI in mental health diagnostics. Prioritizing human-centered and participative design techniques is critical for capturing real users' different requirements and experiences. Bias mitigation strategies, such as inclusive data sourcing and fairness-aware modelling, are critical for preventing algorithmic prejudice. Explainable AI frameworks must be developed to increase physician trust and patient understanding, and interdisciplinary collaboration is essential to ensure that technical systems adhere to clinical, ethical, and societal norms. Furthermore, regulatory bodies and health systems must take proactive steps to build governance structures that protect data rights, ensure informed consent, and provide procedures for recourse in the event of harm. Therefore, the future of AI in mental health diagnosis depends on the advanced technology of its algorithms and the values that define its design and implementation.

References

1. Alanazi HO, Abdullah AH, Qureshi KN. A critical review for developing accurate and dynamic predictive models using machine learning methods in medicine and health care. J Med Syst. 2017;41(4). DOI: 10.1007/s10916-017-0715-6.
2. Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Med Inform Decis Mak. 2020;20(1).
3. Auernhammer J. Human-centered AI: the role of human-centered design research in the development of AI. In:

DRS2020: Synergy. 2020. DOI: 10.21606/drs.2020.282.
4. Carr S. “AI gone mental”: Engagement and ethics in data-driven technology for mental health. J Ment Health. 2020;29(2):1-6. DOI: 10.1080/09638237.2020.1714011.
5. Chen I, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? AMA J Ethics. 2019;21(2):E167-79. DOI: 10.1001/amajethics.2019.167.
6. Chin-Yee B, Upshur R. The impact of artificial intelligence on clinical judgment: a briefing document. Toronto: AMS; 2020. Available from: https://www.ams-inc.on.ca/wp-content/uploads/2020/02/The-Impact-of-AI-on-clinical-judgement.pdf
7. Clark LA, Cuthbert B, Fernández LR, Narrow WE, Reed GM. Three approaches to understanding and classifying mental disorder: ICD-11, DSM-5, and the National Institute of Mental Health’s Research Domain Criteria (RDoC). Psychol Sci Public Interest. 2017;18(2):72-145. DOI: 10.1177/1529100617727266.
8. Colizzi M, Lasalvia A, Ruggeri M. Prevention and early intervention in youth mental health: Is it time for a multidisciplinary and trans-diagnostic model for care? Int J Ment Health Syst. 2020;14(1):1-14. DOI: 10.1186/s13033-020-00356-9.
9. Dousa R. Toward the clinic: understanding patient perspectives on AI and data-sharing for AI-driven oncology drug development. In: IntechOpen. 2020. Available from: https://www.intechopen.com/chapters/72462
10. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health. 2017;4(2). DOI: 10.2196/mental.7785.
11. Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial. JMIR Ment Health. 2018;5(4). DOI: 10.2196/mental.9782.
12. Ceja GE, Riegler M, Nordgreen T, Jakobsen P, Oedegaard KJ, Tørresen J. Mental health monitoring with multimodal sensing and machine learning: A survey. Pervasive Mob Comput. 2018;51:1-26. DOI: 10.1016/j.pmcj.2018.09.003.
13. Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. Artif Intell Healthc. 2020;1(1):295-336. DOI: 10.1016/B978-0-12-818438-7.00012-5.
14. Graham S, Depp C, Lee EE, Nebeker C, Tu X, Kim HC, *et al*. Artificial intelligence for mental health and mental illnesses: an overview. Curr Psychiatry Rep. 2019;21(11):116. DOI: 10.1007/s11920-019-1094-0.
15. Guidi A, Schoentgen J, Bertschy G, Gentili C, Scilingo EP, Vanello N. Features of vocal frequency contour and speech rhythm in bipolar disorder. Biomed Signal Process Control. 2017;37:23-31. DOI: 10.1016/j.bspc.2017.01.017.
16. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. WIREs Data Min Knowl Discov. 2019;9(4). DOI: 10.1002/widm.1312.
17. Jameel T, Ali R, Toheed I. Ethics of artificial intelligence: research challenges and potential solutions. In: 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (ICoMET). 2020. DOI: 10.1109/icomet48670.2020.9073911.
18. Janssen RJ, Miranda MJ, Schnack HG. Making individual prognoses in psychiatry using neuroimaging and machine learning. Biol Psychiatry Cogn Neurosci Neuroimaging. 2018;3(9):798-808. DOI: 10.1016/j.bpsc.2018.04.004.
19. Komeili M, Prom PC, Liaqat D, Fraser KC, Yancheva M, Rudzicz F. Talk2Me: Automated linguistic data collection for personal assessment. PLOS One. 2019;14(3):e0212342. DOI: 10.1371/journal.pone.0212342.
20. Kuziemsky C, Maeder AJ, John O, Gogia SB, Basu A, Meher S, *et al*. Role of artificial intelligence within the telehealth domain. Yearb Med Inform. 2019;28(01):35-40. DOI: 10.1055/s-0039-1677897.
21. Lee NT, Resnick P, Barton G. Algorithmic bias detection and mitigation: best practices and policies to reduce consumer harms. Washington (DC): Brookings Institution; 2019. Available from: https://www.brookings.edu/articles/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/
22. Maher NA, Senders JT, Hulsbergen AFC, Lamba N, Parker M, Onnela JP, *et al*. Passive data collection and use in healthcare: A systematic review of ethical issues. Int J Med Inform. 2019;129:242-7. DOI: 10.1016/j.ijmedinf.2019.06.015.
23. Malik P, Pathania M, Rathaur V. Overview of artificial intelligence in medicine. J Family Med Prim Care. 2019;8(7):2328-31. DOI: 10.4103/jfmpc.jfmpc_440_19.
24. McKinsey & Company. The impact on the workforce and organisations. 2020. Available from: https://eithealth.eu/wp-content/uploads/2020/03/EIT-Health-and-McKinsey_Transforming-Healthcare-with-AI.pdf
25. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncol. 2019;20(5):262-73. DOI: 10.1016/s1470-2045(19)30149-4.
26. Ogundare T. Culture and mental health: towards cultural competence in mental health delivery [Internet]. ResearchGate; 2019 [cited 2025 Oct 30]. Available from: https://www.researchgate.net/publication/338065091_Culture_and_mental_health_Towards_cultural_competence_in_mental_health_delivery
27. Padrez KA, Ungar L, Schwartz HA, Smith RJ, Hill S, Antanavicius T, *et al*. Linking social media and medical record data: A study of adults presenting to an academic, urban emergency department. BMJ Qual Saf. 2015;25(6):414-423. DOI: 10.1136/bmjqs-2015-004489.
28. Soled D. Language and cultural discordance: barriers to improved patient care and understanding. J Patient Exp. 2020;7(6):830-832. DOI: 10.1177/2374373520942398.

29. Stahl BC, Wright D. Ethics and privacy in AI and big data: implementing responsible research and innovation. IEEE Secur Priv. 2018;16(3):26-33.
30. Stangl AL, Earnshaw VA, Logie CH, van Brakel W, Simbayi LC, Barré I, *et al*. The health stigma and discrimination framework: A global, crosscutting framework to inform research, intervention development, and policy on health-related stigmas. BMC Med. 2019;17(1):1-13. DOI: 10.1186/s12916-019-1271-3.
31. Ternes K, Iyengar V, Lavretsky H, Dawson WD, Booi L, Ibanez A, *et al*. Brain health innovation diplomacy: A model binding diverse disciplines to manage the promise and perils of technological innovation. Int Psychogeriatr. 2020;32(8):955-79. DOI: 10.1017/S1041610219002266.