ANALES
DEL SISTEMA SANITARIO DE NAVARRA

An Sist Sanit Navar 2023; 47(2): e1077
https://doi.org/10.23938/ASSN.1077

eISSN: 2340-3527 • ISSN: 1137-6627
https://recyt.fecyt.es/index.php/ASSN/

EDITORIAL

¿Salud para quién? Interseccionalidad y sesgos de la inteligencia artificial para el diagnóstico clínico

Health for whom? Intersectionality and biases in the use of artificial intelligence in clinical diagnosis

Sua Amaya-Santos[1,2,3], **Jaime Jiménez-Pernett**[1,3], **Clara Bermúdez-Tamayo**[1,3,4,5]

Las tecnologías basadas en Inteligencia Artificial (IA) han experimentado en los últimos años un crecimiento significativo en el ámbito de la sanidad, con la promesa de mejorar tanto la atención médica como la salud pública[1]. Así, el potencial evidenciado durante la pandemia por COVID-19 hace que se constituyan como elementos centrales en las estrategias para el desarrollo de sistemas de información sanitaria[2].

Las aplicaciones de la IA se extienden a todo el ciclo de atención al paciente, destacando su habilidad para apoyar el diagnóstico gracias a su capacidad para analizar imágenes médicas como radiografías, resonancias magnéticas y tomografías computarizadas, a través de la identificación de patrones y anomalías que incluso podrían escapar a la detección humana[3].

El uso de estas tecnologías podría conllevar a diagnósticos más tempranos y precisos, así como a minimizar los errores, como han demostrado ya algunas investigaciones en radiología, patología y dermatología que han revelado una mayor precisión diagnóstica al comparar los resultados obtenidos mediante IA con los realizados por profesionales[4]. No obstante, el optimismo hacia los beneficios potenciales de la IA, o *tecno-optimismo*, podría obviar el riesgo de que se perpetúen, exacerben o profundicen los prejuicios y las disparidades en la atención sanitaria por cuestiones de género, raciales o étnicas produciendo inferencias sesgadas. La salud está influida por vínculos complejos entre factores biológicos, socioeconómicos y contextuales, los cuales a menudo se encuentran rodeados de variables de confusión, como el estigma y los estereotipos, que pueden llevar a la representación errónea de los datos[5], y a sesgos cognitivos[6]. Las investigaciones han revelado que los mecanismos de la IA pueden amplificar los comportamientos discriminatorios que son representativos de desigualdades arraigadas[7].

Interseccionalidad en salud

La interseccionalidad constituye un marco esencial para analizar cómo factores como la raza, el género y otras identidades sociales se entrelazan y se combinan, generando manifestaciones de opresión y desigualdad. La idea subyacente es que la investigación científica y la práctica clínica se ha centrado en los miembros más privilegiados, *socavando* los esfuerzos por implementar iniciativas antidiscriminatorias[8].

La evidencia indica que, ante características de las personas, el personal sanitario puede mostrar prejuicios inconscientes que influyen en su interpretación de los síntomas, los resultados de las

1. Escuela Andaluza de Salud Pública. Granada. España.
2. Jagiellonian University of Krakow. Cracovia. Polonia
3. Universidad de Granada. Granada. España.
4. Instituto de Investigación Biosanitaria ibs.GRANADA. Granada. España.
5. CIBER de Epidemiología y Salud Pública (CIBERESP). España.

Correspondencia
Clara Bermúdez-Tamayo [clara.bermudez.easp@juntandeandalucia.es]

Citación:
Amaya-Santos S, Jiménez-Pernett J, Bermúdez-Tamayo C. ¿Salud para quién? Interseccionalidad y sesgos de la inteligencia artificial para el diagnóstico clínico. An Sist Sanit Navar 2024; 47(2): e1077.
https://doi.org/10.23938/ASSN.1077

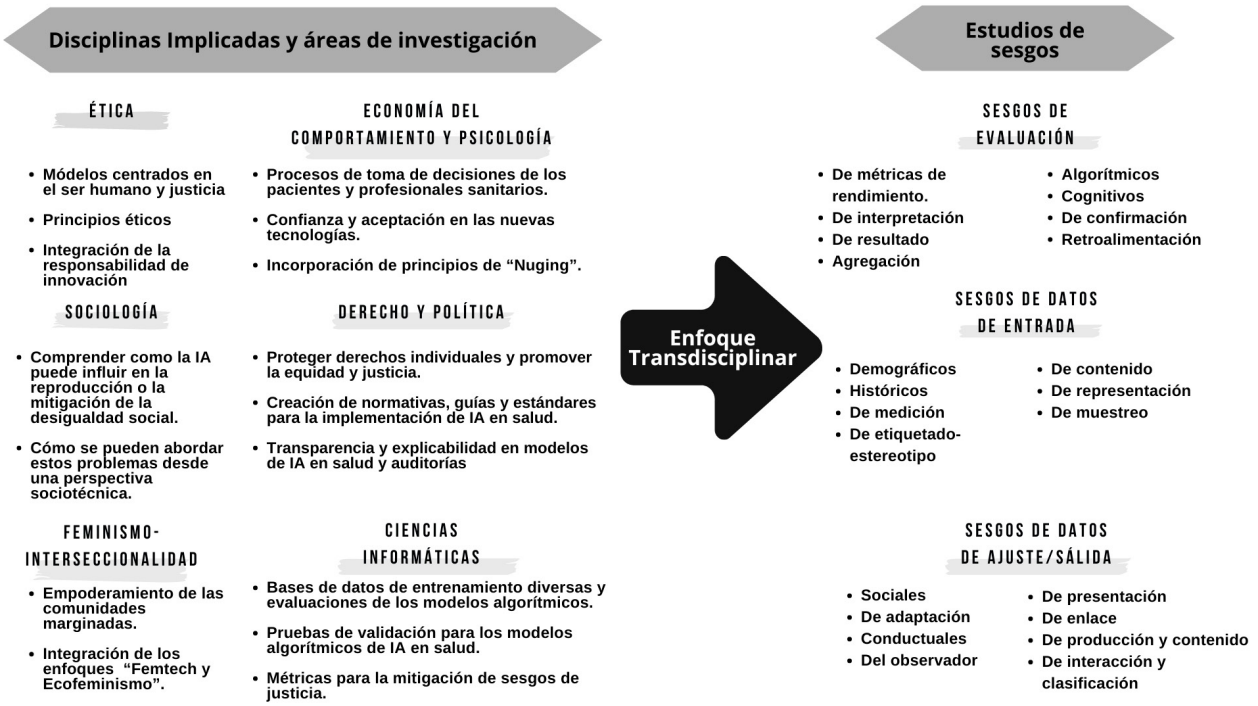

Figura 1. Estudio de sesgos para su mitigación con enfoque transdisciplinar.

pruebas diagnósticas y las recomendaciones de tratamiento[8]. Denominamos *sesgo interseccional* a la discriminación de manera sistemática e injusta contra ciertos individuos o grupos en beneficio de otros, y que puede manifestarse como un sesgo cognitivo, lo que afecta a los procesos de toma de decisiones y dar lugar a disparidades. Es decir, puede conducir a errores diagnósticos, tratamientos subóptimos y daños a los pacientes[9]. Este sesgo podría empeorar la marginación de las minorías y ampliar la brecha de las desigualdades en salud.

Las IA pueden reflejar y amplificar los sesgos presentes en los datos que utilizan, lo que podría perpetuar la discriminación. Este fenómeno se ha abordado desde diversas disciplinas para mitigarlo, dando lugar a diferentes propuestas de clasificación de sesgos. En la figura 1 se describen las disciplinas y los sesgos más frecuentes relacionados con la entrada de datos, algoritmos, evaluación y ajuste/salida[10,11].

Evidencias sobre el sesgo de interseccionalidad en la inteligencia artificial sanitaria

Algunas tecnologías de IA han demostrado ser discriminatorias en función de características como el sexo/género, siendo las mujeres generalmente las más afectadas[12]. Un estudio reciente describe cómo robots entrenados con grandes conjuntos de datos exhibían un comportamiento estereotipado y sesgado en términos de género y raza[13]. Otro estudio mostró sesgo de interseccionalidad del proveedor, documentando los síntomas de pacientes afroamericanos a partir de registros médicos de manera peyorativa[14]. Se ha demostrado que, a partir de imágenes radiológicas, las redes neuronales convolucionales (CNN) pueden subdiagnosticar erróneamente a grupos vulnerables (en particular, hispanos y pacientes con *Medicaid* en Estados Unidos) en una proporción mayor a los pacientes blancos[15]. También se ha evidenciado que un sistema de IA entrenado y validado únicamente en personas con fácil acceso a los servicios produce un sesgo en el diagnóstico a minorías con bajo acceso a la atención sanitaria[14]. Un último caso se refiere al desarrollo de calculadoras de evaluación del riesgo de fractura ósea. Estas calculadoras realizaron correcciones por país para tener en cuenta las diferentes incidencias de osteoporosis (por ejemplo, el riesgo se ajustó a la baja para las mujeres de raza negra con incidencia notificada), pero estas correcciones también generaron infradiagnóstico que sería parte del conjunto de datos que se utilizan para el entrenamiento de IA[16]. Suele asumirse que un aumento en la diversidad cambiará la manera en que se entrenan las IA y, por ende, sus

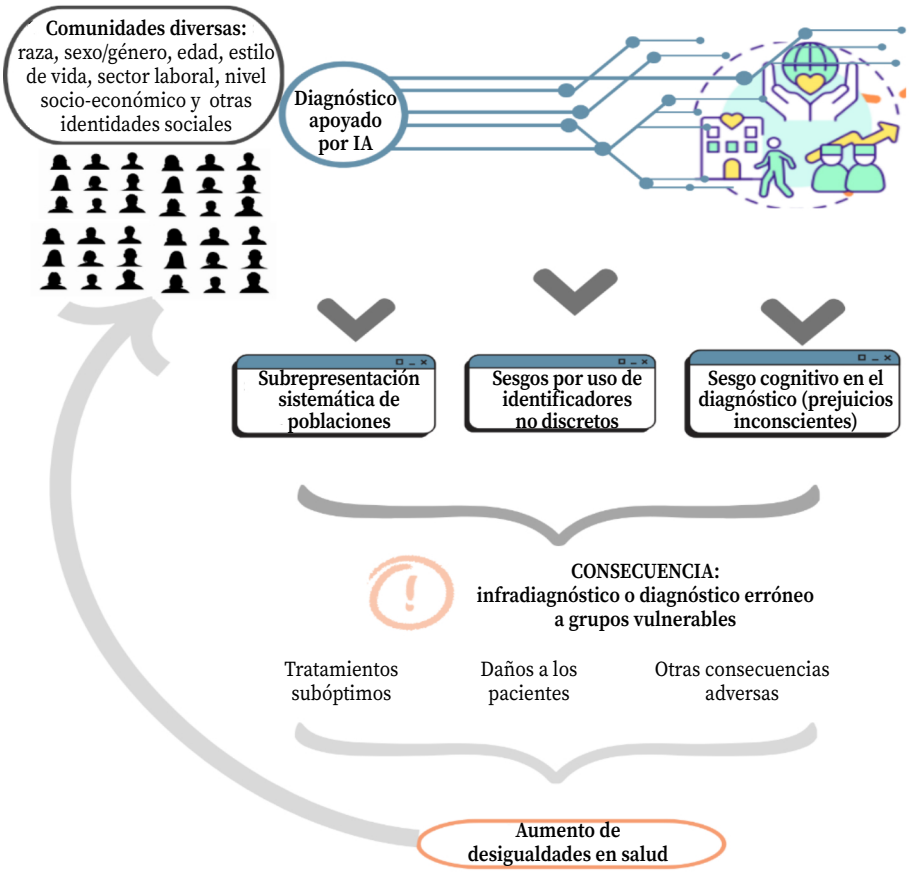

Figura 2. Sesgo de interseccionalidad en el diagnóstico apoyado por Inteligencia Artificial.

predicciones, haciéndola más inclusiva y reduciendo sus riesgos. Sin embargo, esta suposición aún no ha sido probada[9], y lo cierto es que la minimización del sesgo requiere acciones más complejas.

En la figura 2 se esquematiza la interacción de los diferentes mecanismos, y cómo los sesgos de interseccionalidad pueden surgir en el contexto de la IA y sus implicaciones. En primer lugar, la falta de diversidad en los conjuntos de datos digitales usados por los algoritmos de IA puede amplificar la subrepresentación sistemática de ciertas poblaciones[9]. En segundo lugar, los marcadores de identidad pueden ocasionar malentendidos culturales por no recoger la complejidad de los fenómenos de la interseccionalidad (como sexo/género, edad, estilo de vida, o sector laboral) de manera conjunta) estatificándolas y clasificarlos de manera discreta[17] y la asunción de un estatus de hecho imparcial erróneo. Finalmente, hay que considerar el sesgo cognitivo de los profesionales (relacionados con su percepción y razonamiento) implícito en los conjuntos de datos. La toma de decisiones clínicas parece ser el resultado de dos modos distintos de procesamiento cognitivo[18]: el proceso consciente de evaluar opciones basadas en una combinación de utilidad, riesgo, capacidades y/o influencias sociales, o sistema tipo 2, y la cognición automática o sistema tipo 1, referente a los procesos en gran parte inconscientes que ocurren en respuesta a señales ambientales o emotivas y basados en heurísticos arraigados, previamente aprendidos[11,19].

El sesgo de interseccionalidad puede conllevar al infradiagnóstico o diagnóstico erróneo de grupos vulnerables, lo que a su vez podría conllevar a tratamientos subóptimos, daños a los pacientes, amplificar las disparidades de salud y perpetuar la marginalización de grupos desatendidos. Para reconocer y resolver estos problemas es crucial medir el sesgo, tanto en los modelos finales como en los conjuntos de datos, lo que ha llevado al desarrollo de métricas para la detección de sesgos en los últimos años. Así, un estudio empleó diferentes taxonomías de métricas de sesgo demográfico para detectar el sesgo representacional y estereotípico en bases de datos para el entrenamiento de una IA para el reconocimiento de expresiones faciales[19].

Políticas, marcos y directrices en materia de IA

La normativa desempeña un papel central en el establecimiento de un marco defensivo frente a las amenazas percibidas, anticipadas y reales de la IA[21]. Los esfuerzos para abordar sus riesgos e implicaciones sociales y éticas han dado lugar a un *corpus* documental cada vez más extenso, dado que la mayoría de los instrumentos regulatorios existentes no fueron redactados teniendo en cuenta la magnitud de los cambios de la IA[22]. Distintos países avanzan en el abordaje de estas brechas. Por ejemplo, la Comisión Europea propuso en 2021 una Ley de Inteligencia Artificial, actualmente en desarrollo, aplicable a los sistemas de IA en salud, aunque por ahora no aborda suficientemente las especificidades de este campo[23].

Otras iniciativas se refieren a retos en materia de responsabilidad que requieren una atención política urgente, como derechos humanos, cuestiones sociales, económicas y medioambientales, y valores democráticos, integrados bajo el nombre de *Inteligencia artificial responsable* (IAR)[24] (Fig. 3). Iniciativas tales como los *Lineamientos para la Inteligencia Artificial Responsable* o la herramienta de evaluación *Responsabilidad de las soluciones digitales en salud* impulsadas recientemente por compañías de *Big Tech*, el mundo académico e institutos de investigación, junto con gobiernos y ONG. Desafortunadamente, hasta ahora ha tenido escaso impacto en la práctica real de la IA[25].

Desafíos para el futuro

Abordar la cuestión de *¿Salud para quién?* requiere ir más allá de directrices e intenciones, con normativas que prioricen el desarrollo y el uso responsable de la IA centrándose en el bienestar de todos los individuos, para que los responsables políticos refuercen su papel en un campo tan dinámico.

La confiabilidad de la IA es un cuello de botella crítico en su adopción. El fenómeno de la *caja negra* es una crítica a la mayoría de la IA actual, ya que carece de transparencia, de explicación, y sus resultados no pueden generalizarse[26]. Una sugerencia general es aumentar la diversidad entre el personal investigador de distintas disciplinas que trabajan con macrodatos/IA y equidad. Los sesgos discriminatorios se pueden prevenir mediante la incorporación de una amplia gama de perspectivas, ya que esto puede reducir la probabilidad de generar sesgos basados en puntos de vista singulares[27].

Se necesita más investigación para determinar la mejor manera de detectar el sesgo relacionado con la interseccionalidad por el uso de identificadores no discretos[16]. Esto implica que las políticas nacionales, las instituciones y las comunidades de investigación tendrían que profundizar en el desarrollo de estándares armonizados en interseccionalidad e identidades sociales[9,17].

Los marcos legales existentes tienden a poner un énfasis en la seguridad física y la privacidad, descuidando factores igualmente importantes como

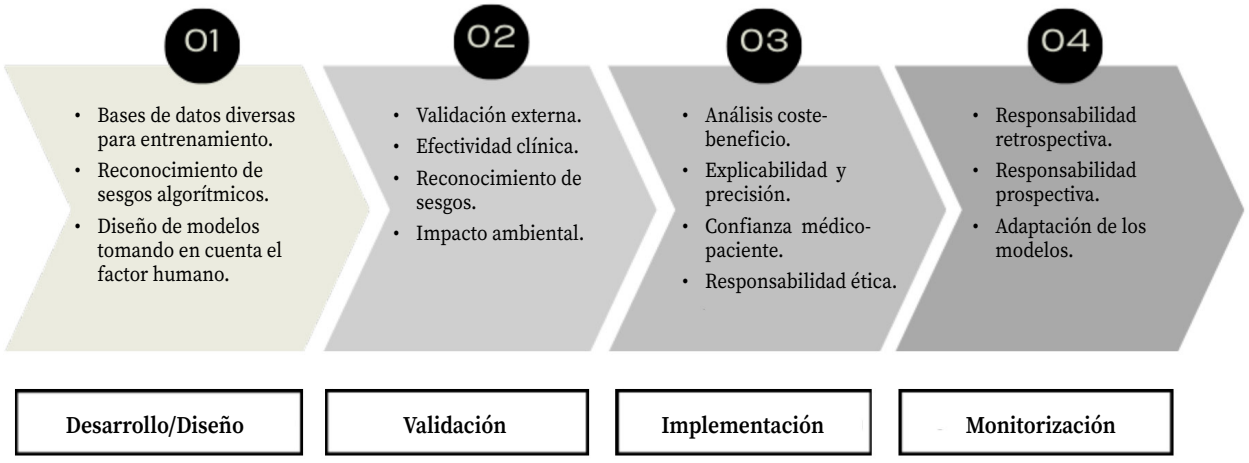

Figura 3. Proceso para una Inteligencia Artificial responsable en salud. Elaboración propia a partir de Sujan y col 2023[29].

la diversidad, la subrepresentación sistemática de poblaciones y la influencia del sesgo cognitivo implícito en los datos[28]. Es necesario, además, considerar tanto la responsabilidad retrospectiva como la responsabilidad prospectiva. La primera implica rendir cuentas y/o la necesidad de poder comprender y explicar las decisiones de dichos sistemas. Por otra parte, la responsabilidad prospectiva exige que todas las partes interesadas asuman el deber de garantizar un despliegue ético de la IA[29].

BIBLIOGRAFÍA

1. Bermudez-Tamayo C, Jimenez-Pernett J. Inteligencia artificial para el avance de los sistemas de salud. Posibles aportes y retos. Revista de Derecho de la Seguridad Social Laborum 2022; 4(especial): 401-414. https://revista.laborum.es/index.php/revsegsoc/article/view/641/735
2. Peiró, S. El futuro de la salud pública tras la pandemia. Una ventana de oportunidad. An Sist Sanit Navar 2023; 46(2): e1045. https://doi.org/10.23938/ASSN.1045
3. Murphy K, Di Ruggiero E, Upshur R, Willison D, Malhotra R, Cai J. Artificial intelligence for good health: a scoping review of the ethics literature. BMC Med Ethics 2021; 22(1):14. https://doi.org/10.1186/s12910-021-00577-8
4. Miller, D, Brown E. Artificial intelligence in medical practice: The question to the answer? Am J Med 2018; 131(2): 129-133. https://doi.org/10.1016/j.amjmed.2017.10.035
5. Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. NPJ Digit Med 2020; 3: 81. https://doi.org/10.1038/s41746-020-0288-5
6. Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artif Intell 2019; 267: 1-38. https://doi.org/10.1016/j.artint.2018.07.007
7. Palacios Barea MA, Boeren D, Ferreira Goncalves JF. At the intersection of humanity and technology: a technofeminist intersectional critical discourse analysis of gender and race biases in the natural language processing model GPT-3. AI & Soc 2023. https://doi.org/10.1007/s00146-023-01804-z
8. Wilson Y, White A, Jefferson A, Danis M. Intersectionality in clinical medicine: The need for a conceptual framework. Am J Bioeth 2019; 19(2): 8-19. https://doi.org/10.1080/15265161.2018.1557275
9. Nasir S, Khan RA, Bai S. Ethical framework for harnessing the power of AI in healthcare and beyond. IEEE Access 2023; 12: 31014–31035. https://doi.org/10.1109/ACCESS.2024.3369912
10. Ashokan A, Haas C. Fairness metrics and bias mitigation strategies for rating predictions. Information Processing & Management 2021; 58(5): 102646. https://doi.org/10.1016/j.ipm.2021.102646
11. García Mochon L, Olry de Labry Lima A, Bermudez Tamayo C. Prioritization of non-recommended clinical activities in Primary Care. An Sist Sanit Navar 2017; 40(3): 401-412. https://doi.org/10.23938/ASSN.0120
12. Ciston S. Intersectional artificial intelligence is essential: Polyvocal, multimodal, experimental methods to save AI. J Sci Technol Arts 2019; 11(2): 3-8. https://doi.org/10.7559/citarj.v11i2.665
13. O'Connor S, Liu H. Gender bias perpetuation and mitigation in AI technologies: challenges and opportunities. AI & Soc 2023. https://doi.org/10.1007/s00146-023-01675-4
14. Buslón N, Cortés A, Catuara-Solarz S, Cirillo D, Rementeria MJ. Raising awareness of sex and gender bias in artificial intelligence and health. Front Glob Womens Health 2023; 4: 970312. https://doi.org/10.3389/fgwh.2023.970312
15. Abràmoff MD, Tarver ME, Loyo-Berrios N, Trujillo S, Char D, Obermeyer Z et al. Considerations for addressing bias in artificial intelligence for health equity. NPJ Digit Med 2023; 6: 170. https://doi.org/10.1038/s41746-023-00913-9
16. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 2021; 27(12): 2176-2182. https://doi.org/10.1038/s41591-021-01595-0
17. Lewiecki EM, Wright NC, Singer AJ. Racial disparities, FRAX, and the care of patients with osteoporosis. Osteoporos Int 2020; 31, 2069-2071. https://doi.org/10.1007/s00198-020-05655-y
18. Kahneman D. Thinking, Fast and Slow. New York: Farrar, Straus and Giroux, 2011.
19. Dominguez-Catena I, Paternain D, Galar M. Metrics for dataset demographic bias: A case study on facial expression recognition. IEEE Trans Pattern Anal Mach Intell 2024; 1-18. https://doi.org/10.1109/TPAMI.2024.3361979
20. Bermúdez-Tamayo C, Olry de Labry-Lima A, García- Mochón L. No hacer: de las recomendaciones a la acción. An Sist Sanit Navar 2019; 42(1): 105-108. https://doi.org/10.23938/ASSN.0380
21. Thomasian NM, Eickhoff C, Adashi EY. Advancing health equity with artificial intelligence. J Public Health Pol 2021; 42(4): 602-611. https://doi.org/10.1057/s41271-021-00319-5
22. Roche C, Wall PJ, Lewis D. Ethics and diversity in artificial intelligence policies, strategies and initiatives. AI Ethics 2022; 3: 1095-1115. https://doi.org/10.1007/s43681-022-00218-9
23. Mezgár I, Váncza J. From ethics to standards: A path via responsible AI to cyber-physical production systems. An Rev Control 2022; 53: 391-404. https://doi.org/10.1016/j.arcontrol.2022.04.002

24. Lehoux P, Rivard L, de Oliveira RR, Mörch CM, Alami H. Tools to foster responsibility in digital solutions that operate with or without artificial intelligence: A scoping review for health and innovation policymakers. Int J Med Inform 2023; 170: 104933. https://doi.org/10.1016/j.ijmedinf.2022.104933
25. Organization for Economic Co-operation and Development. The state of implementation of the OECD AI Principles four years on. OECD Artificial Intelligence Papers No. 3. Paris: OECD Publishing, 2023. https://doi.org/10.1787/dee339a8-en
26. Mollura DJ, Culp MP, Pollack E, Battino G, Scheel JR, Mango VL et al. artificial intelligence in low- and middle-income countries: Innovating global health radiology. Radiology 2020; 297(3): 513-520. https://doi.org/10.1148/radiol.2020201434
27. Wesson P, Hswen Y, Valdes G, Stojonowski K, Handley MA. Risks and opportunities to ensure equity in the application of big data research in public health. Annu Rev Public Health 2022; 43: 59-78. https://doi.org/10.1146/annurev-publhealth-051920-110928
28. Obasa AE, Palk AC. Responsible application of artificial intelligence in health care. South African Journal of Science 2023; 119(5-6): 1-3.
29. Sujan M, Smith-Frazer C, Malamateniou C, Connor J, Gardner A, Unsworth H et al. Validation framework for the use of AI in healthcare: overview of the new British standard BS30440. BMJ Health Care Inform 2023; 30(1): e100749. https://doi.org/10.1136/bmjhci-2023-100749