TELEHEALTH AND MEDICINE TODAY

DEBATE/COMENTARIOS/PERSPECTIVAS

El papel de la IA explicable y los marcos de evaluación para la integración segura y eficaz de grandes modelos lingüísticos en la atención sanitaria

Sandeep Reddy, MBBS, MSc, PhD[1] ; Alexandre Lebrun, MS, MSEE[2] ; Adam Chee, PhD[3] ; y Dimitrios Kalogeropoulos, PhD[4,5]

[1]Director, Facultad de Medicina Deakin, Victoria (Australia); [2]Cofundador y director general, Nabla, París (Francia); [3]Profesor asociado, Saw Swee Hock (Reino Unido).
Saw Swee Hock School of Public Health, National University of Singapore, Singapur; [4]) Fundador y Director Ejecutivo, Global Health & Digital Innovation Foundation, Londres, Reino Unido; [5]UCL Global Business School for Health, Londres, Reino Unido.

Autor correspondiente: Sandeep Reddy; Correo electrónico: sandeep.reddy@deakin.edu.au

DOI: https://doi.org/10.30953/thmt.v9.485

Palabras clave: IA, inteligencia artificial, Converge2Xcelerate, sanidad, grandes modelos lingüísticos, LLM

Enviado: 8 de febrero de 2024; Aceptado: 2 de abril de 2024; Publicado: 30 de abril de 2024

Ta integración de la inteligencia artificial (IA), en particular los grandes modelos lingüísticos (LLM, por sus siglas en inglés: un tipo especializado de IA que se entrena con grandes cantidades de texto), en las tecnologías de la información y la comunicación (TIC). para comprender el contenido existente y generar contenido original) en la atención sanitaria sigue acelerándose, lo que exige una evaluación y supervisión minuciosas para garantizar un despliegue seguro, ético y eficaz. Durante una mesa redonda celebrada en el evento 2023 Converge2Xcelerate, organizado por la editorial de esta revista, Partners in Digital Health, expertos en IA compartieron perspectivas destacadas sobre las oportunidades y los retos de trasladar con sensatez innovaciones como LLM a la atención sanitaria.(1)A continuación se presenta un resumen de las perspectivas clave de la mesa redonda.

Principales temas tratados:

- El potencial de la IA explicable para facilitar la transparencia y la confianza;
- Desafíos a la hora de alinear la IA con protocolos sanitarios globales variables;
- La importancia de la evaluación a través de marcos traslacionales y de gobernanza adaptados a los contextos sanitarios;
- Escepticismo en torno a los usos excesivamente amplios de los LLM para interfaces conversacionales;
- La necesidad de validar juiciosamente los LLM, teniendo en cuenta los niveles de riesgo.

Además, en el debate se destacó la explicabilidad (es decir, el concepto de que un modelo de aprendizaje automático y sus resultados pueden explicarse y "tienen sentido").

Además, se destacó la explicabilidad (es decir, el concepto de que un modelo de aprendizaje automático y sus resultados pueden explicarse y "tener sentido" para un ser humano a un nivel aceptable), la evaluación y la deliberación cuidadosa con los profesionales sanitarios como elementos fundamentales para obtener beneficios y, al mismo tiempo, abordar de forma proactiva los riesgos de una mayor adopción de la IA en medicina. Otros temas de debate se centraron en las prácticas de evaluación crítica de la IA en medicina, la necesidad y las limitaciones de la IA explicable, y las deliberaciones que los responsables sanitarios deben llevar a cabo antes de su implantación.

La adopción de técnicas de Inteligencia Artificial (IA), en especial LLMs como GPT-4 (Generative Pre-trained Transformer 4), en la administración sanitaria y la práctica clínica se está acelerando rápidamente(2).Aunque promete aumentar las capacidades humanas y mejorar el acceso, la calidad y la eficiencia, la integración de la IA introduce complejas consideraciones técnicas, éticas y normativas relativas a la transparencia, además de la responsabilidad y el impacto en los pacientes y los profesionales sanitarios(3,4).

IA explicable: generar confianza y transparencia

Un punto central de la conversación fue el papel de las técnicas de IA explicable a la hora de establecer la confianza en los sistemas de IA. Los participantes coincidieron en que, aunque los LLM pueden proporcionar resultados útiles, funcionan intrínsecamente como "cajas negras" que ocultan el razonamiento que subyace a las conclusiones. Esta opacidad resulta preocupante en contextos sanitarios de alto riesgo.

Los panelistas destacaron la IA explicable como un área activa de investigación para abordar estos problemas de transparencia.

Sin embargo, lograr una explicabilidad completa con redes neuronales de gran tamaño sigue siendo un reto. La naturaleza compleja, multivariable y dinámica de los entornos clínicos puede confundir aún más los enfoques de explicación adaptados a entornos más restringidos.[6]

Evaluación: Garantizar una traducción eficaz y ética

Estas observaciones subrayan la necesidad de una evaluación rigurosa adaptada a las aplicaciones sanitarias a lo largo de todo el diseño, despliegue y funcionamiento de los sistemas de IA. Los ponentes abogaron por marcos de evaluación amplios y continuos, como TEHAI (Translational Evaluation of Healthcare AI)[7],que abarcan la integración con los sistemas informáticos de asistencia sanitaria, los factores de adopción clínica, la supervisión actualizada del rendimiento y los componentes de gobernanza[8],como la rendición de cuentas y la ética. Los panelistas también señalaron las diferencias con respecto a las evaluaciones reglamentarias más limitadas de la seguridad y el daño por sí solas. La evaluación multidimensional puede poner de manifiesto los puntos fuertes y débiles, así como los casos de uso apropiados para orientar la adopción de la IA de forma responsable.

Automatizar los procesos asistenciales, no las conversaciones

Otra consideración que se planteó durante la mesa redonda fue la creciente propensión a antropomorfizar los sistemas de IA, especialmente las interfaces conversacionales. Los panelistas advirtieron de que una personificación excesiva genera expectativas poco prácticas en torno a la exhaustividad y la perfección. Sí, los LLM pueden ayudar a automatizar procesos asistenciales clave, pero actualmente el diagnóstico o el tratamiento totalmente automatizados siguen siendo poco fiables.

Además, los panelistas sugirieron centrar la automación en el apoyo a los médicos en las tareas de documentación, codificación y coordinación asistencial. Las explicaciones sobre la inteligencia artificial sirven para construir modelos mentales adecuados sobre sus capacidades y limitaciones más que para humanizarla.[9]

Avanzar con prudencia

En resumen, los participantes en la mesa redonda destacaron que los enfoques cautelosos, colaborativos[8]y evaluativos son imprescindibles para aplicar de forma productiva los avances de la IA en la práctica clínica. Más que en la sustitución, se hace hincapié en empezar con funciones de apoyo y menor riesgo, y en evaluar de forma transparente las ventajas y los inconvenientes revelados por la aplicación en el mundo real.[7]

El progreso continuado depende de evaluaciones abiertas y honestas por parte de líderes interdisciplinares para encontrar los nichos más constructivos para innovaciones emergentes como los LLM en la prestación real de cuidados. Los marcos de evaluación explicables que facilitan la transparencia pueden desempeñar un papel integral en el avance responsable de la integración de la IA.

Financiación

En la preparación de este manuscrito no se ha utilizado financiación externa.

Relaciones y actividades financieras y no financieras

Sandeep Reddy preside y es director de Healea Services y del Centre for Advancement of Translational AI in Medicine.

Colaboradores

Sandeep Reddy se encargó de redactar y finalizar el manuscrito. Todos los autores han contribuido al contenido del manuscrito.

Declaración de disponibilidad de datos (DAS), intercambio de datos, reproducibilidad y repositorios de datos

En la preparación de este artículo no se han utilizado datos originales.

Aplicación de texto generado por IA o tecnología relacionada

Se utilizó tecnología de IA para generar la transcripción de la mesa redonda y ayudar con la gramática en la preparación del manuscrito.

Agradecimientos

Los autores agradecen a Partners in Digital Health la organización del evento y de la mesa redonda.

Referencias

1. Telesalud y medicina hoy. Simposio ConV2X [Internet]. [citado 2024 Feb 1]. Disponible en: https://telehealth.conv2xsym-posium.com
2. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial Intelligence (AI) and global health: how can AI con-tribute to health in resource-poor settings? BMJ Glob. Health. 2018;3:e000798. https://doi.org/10.1136/bmjgh-2018-000798
3. Char DS, Shah NH, Magnus D. Implementing machine learn-ing in health care-addressing ethical challenges. N Engl J Med. 2018;378:981–3. https://doi.org/10.1056/NEJMp1714229
4. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Desafíos clave para entregar impacto clínico con inteligencia artifi-cial. BMC Med. 2019;17. https://doi.org/10.1186/ s12916-019-1426-2
5. Adadi A, Berrada M. Peeking inside the black-box: a sur-vey on explainable artificial intelligence (XAI). IEEE Access. 2018;6:52138–60. https://doi.org/10.1109/ACCESS.2018.2870052
6. Rudin C. Stop explaining black box machine learning mod-els for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1:206-15. https://doi.org/10.1038/ s42256-019-0048-x
7. Reddy S, Rogers W, Makinen VP, Coiera E, Brown P, Wenzel M, et al. Evaluation framework to guide implementation of AI systems into healthcare settings. BMJ Health Care Inform. 2021;28(1):e100444. https://doi.org/10.1136/bmjhci-2021-100444

8. Kalogeropoulos D, Barach P. Telehealth's role enabling sustainable innovation and circular economies in health. Telehealth Med Today. 2023;8(1). https://doi.org/10.30953/thmt.v8.409
9. 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