CONSENSO DE SALUD DIGITAL
SOCIEDAD ARGENTINA DE CARDIOLOGÍA ABRIL 2026 PRIMERA PARTE

Directores
Sebastián S. Benitez MTSAC
Mariano Benzadón MTSAC

Sub director
Federico Cintora MTSAC
Secretario
Gabriel Perea MTSAC
Coordinadores
Gustavo Daquarti
Eduardo Filipini MTSAC
Rocio Blanco
Ignacio Ríos MTSAC
Jorgelina Medus MTSAC
Marina Baglioni MTSAC
Agustina Miragaya

Comité de redacción
Juan Martín Adjigogovic
Sebastián Aguiar
Franco Ballari
Maico Bernal
Juan Pablo Bonifacio
Guillermo Borel
Carlos David Bruque
Lucrecia Burgos MTSAC
Daniela Cardillo
Adrián Carlessi
Flavio Colavecchia
Pablo Elissamburu MTSAC
Mariel Foti
Enrique Gómez MTSAC
Mariana Guerchicoff
Neiva Maciel MTSAC
Silvina Manuela Mansilla
German Matto
Lucas Müller
Sebastián Obregón MTSAC
Carlos Otero
Lucas Suárez
Juan Manuel Rodríguez Platas
Nicolás Vecchio
Victoria Saez
Daniel Zaffora

Revisores
Silvia Makhoul MTSAC
María Luz Fernández Recalde MTSAC
Por ACN
Guadalupe Pagano

INTRODUCCIÓN

¿Por qué un consenso de salud digital en cardiología? Quizás la mejor respuesta sea otra pregunta, ¿por qué la cardiología o la medicina en general no debería ponerse de acuerdo en cómo adaptarse e implementar las evoluciones continuas y constantes de las tecnologías que cambian radicalmente nuestro día a día?

Por lo cual comenzaremos por unificar y recomendar que todo lo referido a innovación y usos de tecnologías de la información y comunicación (TICs) aplicadas en el ámbito de la salud sean denominados bajo el término de Salud Digital. Por lo que definiremos a la Salud Digital como el uso de las TICs en los procesos de atención sanitaria, éstas deben ser empleadas de manera complementaria e innovadora para tratar las necesidades de salud de la población; es decir, nuestra práctica en algunas ocasiones será en el consultorio, en otras en la sala de internación o unidad coronaria y por momentos la misma será a través de las TICs.

Cuando se habla de TICs, se puede referir a diferentes criterios según el contexto de uso del término. A saber, redes, refiriéndose tanto a las redes de radio y televisión, como a las redes de telefonía fija y móvil, así como el ancho de banda. Terminales y equipos: abarca aparatos mediante los que operan las redes de información y comunicación, por ejemplo: computadoras, tabletas, teléfonos celulares, dispositivos de audio y vídeo, televisores, consolas de juego. Servicios, se refiere al amplio espectro de servicios ofrecidos con recursos anteriores; por ejemplo: servicios de correo electrónico, almacenamiento en la nube, educación a distancia, banca electrónica, juegos en línea, servicios de entretenimiento, comunidades virtuales, blogs e inteligencia artificial entre otros.

Es indiscutible que hay elementos del acto médico que hasta ahora sólo se logran de manera presencial, como es el examen físico, que requiere de la interacción del profesional con el paciente en forma personal, en el ámbito adecuado y respetando todas las implicancias ético-legales que rigen la profesión. Mientras esta actividad no sea reconocida como acto médico pleno e idóneo (para determinadas intervenciones) será muy difícil conseguir la jerarquización y regulación de honorarios profesionales para la misma.

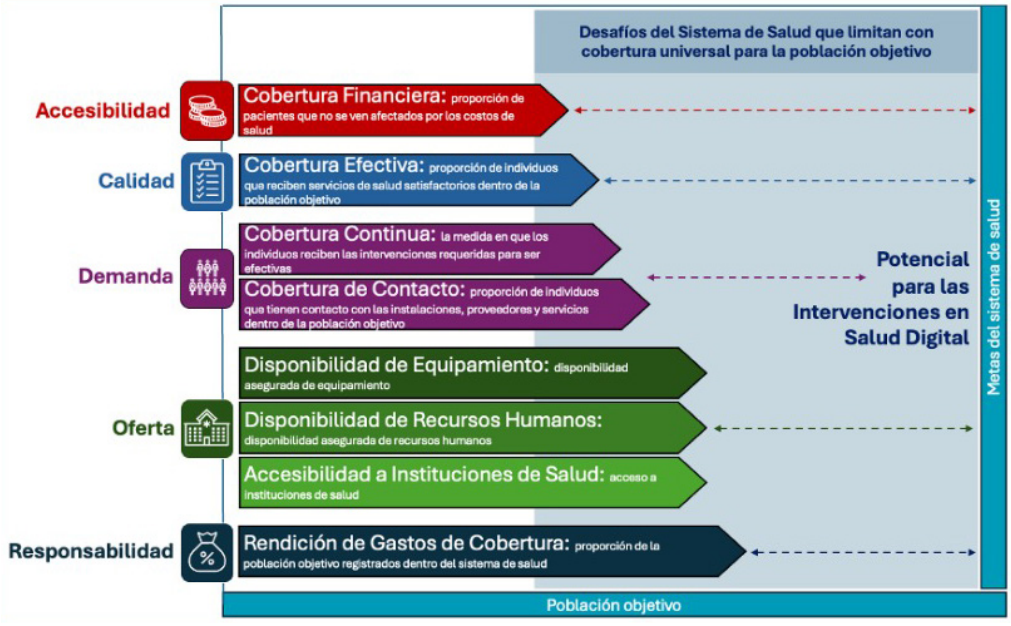
Rol de la salud digital en el fortalecimiento del sistema de salud

El ideal de cobertura de salud tiene como objetivo garantizar la calidad, la accesibilidad y la disponibilidad de los servicios de salud. Sin embargo, sigue habiendo deficiencias para garantizar el acceso a todos los que necesitan servicios de salud y para garantizar que se presten con la calidad prevista sin causar dificultades financieras.

El modelo Tanahashi, publicado por la OMS en 1978, proporciona un marco probado en el tiempo para comprender las brechas de los sistemas de salud y cómo estas dificultan la cobertura, la calidad y la disponibilidad previstas de los servicios de salud para las personas. Este modelo en cascada ilustra cómo los sistemas de salud pierden rendimiento debido a los desafíos en los niveles sucesivos, cada uno de los cuales depende del nivel anterior. Los desafíos del sistema de salud, como la inaccesibilidad geográfica, la baja demanda de servicios, el retraso en la prestación de atención, la baja adherencia a los protocolos clínicos y los costos para las personas y los pacientes, contribuyen a pérdidas incrementales en el desempeño del sistema de salud que repercuten acumulativamente en la salud de las personas. Estos déficits limitan la capacidad de cerrar las brechas y socavan el potencial para lograr la mejor cobertura posible.

En el mismo año, por medio de su Declaración de Alma Ata la OMS define la atención primaria de la salud como la atención de salud esencial basada en métodos y tecnologías prácticos, científicamente sólidos y socialmente aceptables, que se hacen universalmente accesibles a las personas y las familias de la comunidad mediante su plena participación y a un costo que la comunidad y el país pueden permitirse mantener en todas las etapas de su desarrollo en un espíritu de autosuficiencia y libre determinación. Es el primer nivel de contacto de las personas, la familia y la comunidad con el sistema nacional de salud, acercando la atención de salud lo más posible al lugar donde las personas viven y trabajan, y constituye el primer elemento de un proceso de atención de salud continuado. Esta declaración de avanzada hasta la fecha se ha visto limitada en su implementación, con muchas brechas para su cumplimiento. Casi 50 años después, a través de las intervenciones en salud digital, puede que la atención primaria de la salud cuente con las herramientas adecuadas para saltar estas brechas y lograr la consecución de su fin.

Fig. Modelo Tanahashi adaptado sobre el desempeño del sistema de salud.



INTERVENCIONES EN SALUD DIGITAL

Podemos definir intervenciones de salud digitales como una tecnología digital que es utilizada para alcanzar los objetivos en salud, pudiendo ser un punto de partida para categorizar las que se utilizan para superar los desafíos definidos del sistema de salud. A los fines del presente consenso se entiende por intervenciones en salud digital a:

Teleconsulta: interacción donde el paciente consulta directamente al profesional utilizando la vía de la tecnología digital.

Teleinterconsulta: se define a la teleinterconsulta como la interacción entre dos médicos. Uno de los cuales es denominado consultor y puede encontrarse físicamente presente con el paciente o aún en ausencia de este.

El otro médico que es quien se desempeña de manera remota (a la distancia, ya sea se encuentre en territorio nacional o en el extranjero) es usualmente reconocido por su especialidad o por ser de alguna manera destacado en el manejo de un problema médico en particular siendo por ello que se lo requiere especialmente, siendo esto responsabilidad del consultor.

Telemonitoreo: el monitoreo constituye un acto profesional médico, que integra la atención del paciente, la cual importa su seguimiento en el tiempo. El telemonitoreo utiliza este sistema de monitoreo de parámetros y métodos complementarios, siendo útiles para el diagnóstico, evaluación de tratamientos y seguimiento del paciente pero que son enviados al profesional de la salud utilizando la vía de la tecnología digital.

Telegestión: la misma comprende la comunicación digital para el desarrollo e implementación de: documentales médicas y registros, extensiones de recetas y/o prescripciones de estudios, almacenamiento y transferencia de métodos complementarios de diagnósticos, transmisión de notificaciones y flujos de trabajo a los profesionales de la salud.

Teleeducación: alcanza a las diferentes conferencias, cursos y debates entre especialistas (expertos en temas específicos) transmitidas por la vía de la tecnología digital a profesionales de todo el país y también desde y hacia el extranjero, como el uso de las tecnologías de la información y comunicación como herramientas de soporte del conocimiento médico (por ejemplo tarjetas de ayuda cognitiva) y para la transmisión de información segura y de fuente verificable al paciente en particular como a la población en general.

Recomendaciones

- Se definen como intervenciones en salud digital a la teleconsulta, teleinterconsulta, telemonitoreo, telegestión y teleeducación. Recomendación de Clase I, Nivel de evidencia C.
- Las intervenciones en salud digital deben ser consideradas como un acto médico pleno e idóneo. Recomendación de Clase I, Nivel de evidencia C.
- Las intervenciones en salud digital, como todo otro acto médico, conlleva la contraprestación del pago de la práctica u honorarios correspondiente. Recomendación de Clase I, Nivel de evidencia C.

BIBLIOGRAFÍA

1. Organización Mundial de la Salud (2019). WHO guideline: recommendations on digital intervention for health system strengthening. ISBN 978-92-4-155050-5.
2. Organización Mundial de la Salud (1978). Informe de la Conferencia Internacional sobre Atención Primaria de la Salud. Alma-Ata, URSS.ISBN 92-4-354135-8.
3. Tromp J, Jindal D, Redfern J, Bhatt A, Séverin T, Banerjee A, et al. World Heart Federation Roadmap for Digital Health in Cardiology. Glob Heart 2022;17:61.https://doi.org/10.5334/gh.1141

MÉTODOS

Clases de recomendación y niveles de evidencia

Se detallan el grado de consenso y los niveles de evidencia alcanzados para establecer las recomendaciones finales que se basaron en el reglamento del Área de Normatización y Consensos de la Sociedad Argentina de Cardiología.

Clases de recomendación

- Clase I: condiciones para las cuales hay evidencia y/o acuerdo general en que el tratamiento/procedimiento es beneficioso, útil y eficaz.

- Clase II: evidencia conflictiva y/o divergencia de opinión acerca de la utilidad, eficacia del método, procedimiento y/o tratamiento.

• Clase IIa: el peso de la evidencia/opinión está a favor de la utilidad/eficacia.

• Clase IIb: la utilidad/eficacia está menos establecida.

- **Clase III:** evidencia o acuerdo general de que el tratamiento método/procedimiento no es útil/eficaz y en algunos casos puede ser perjudicial.

Niveles de evidencia

- Nivel de evidencia A: evidencia sólida, proveniente de estudios clínicos aleatorizados o de cohortes con diseño adecuado para alcanzar conclusiones estadísticamente correctas y biológicamente significativas.

- Nivel de evidencia B: datos procedentes de un único ensayo clínico aleatorizado o de grandes estudios no aleatorizados.

- Nivel de evidencia C: consenso de opinión de expertos.

Identificación de temas prioritarios

Se consideran en el marco de este consenso como temas prioritarios la equidad, la eficiencia, no discriminación, la universalidad, la seguridad, la calidad de la atención de salud, la solidaridad, la preservación de la relación médico - paciente, en el marco del respeto a la confidencialidad y secreto médico, la accesibilidad, la educación y el respeto a la personalidad, la dignidad humana y la identidad, de acuerdo con las disposiciones legales vigentes.

La salud digital debe entenderse como un medio complementario, facilitador y no sustitutivo de la atención presencial, en un ámbito más donde el profesional desarrolla su tarea. El uso de estas herramientas en todos los casos debe realizarse contemplando el mejor interés de la población (pacientes, equipo de salud, instituciones de la población civil y del estado).

También es prioridad promover las capacitaciones necesarias de los recursos humanos en salud para el empleo de las herramientas de salud digital en la atención de los usuarios de los servicios. Articular con otras instituciones científicas y con los niveles provinciales y nacional para el desarrollo de un sistema de colaboración que posibilite el crecimiento y la comunidad en materia de innovación e implementación de las intervenciones en salud digital.

BIBLIOGRAFÍA

1. Organización Mundial de la Salud (2019). WHO guideline: recommendations on digital intervention for health system strengthening. ISBN 978-92-4-155050-5.
2. Organización Mundial de la Salud (1978). Informe de la Conferencia Internacional sobre Atención Primaria de la Salud. Alma-Ata, URSS. ISBN 92-4-354135-8.
3. Omboni S. Connected Health in Hypertension Management. Front Cardiovasc Med 2019;6:76. https://doi.org/10.3389/fcvm.2019.00076
4. Tromp J, Jindal D, Redfern J, Bhatt A, Séverin T, Banerjee A, et al. World Heart Federation Roadmap for Digital Health in Cardiology. Glob Heart 2022;17:61.https://

Desarrollo

Inteligencia artificial

Definiciones relacionadas al capítulo de Inteligencia Artificial

La incorporación de inteligencia artificial (IA) en el ámbito de la salud exige una comprensión compartida de los términos fundamentales que estructuran esta disciplina. Contar con definiciones claras permite alinear criterios, mejorar la comunicación entre equipos interdisciplinarios y establecer un marco común para la adopción segura, ética y efectiva de estas tecnologías.

Inteligencia Artificial (IA): Es la capacidad de una máquina para realizar tareas que

normalmente requieren inteligencia humana, como razonar, aprender o tomar decisiones.

Machine Learning: Es una técnica dentro de la inteligencia artificial que permite que las computadoras aprendan automáticamente de la experiencia (datos) y mejoren su desempeño sin necesidad de ser programadas paso a paso.

Deep Learning: Es una forma avanzada de machine learning que utiliza algoritmos basados en redes neuronales artificiales de múltiples capas, inspiradas en la estructura del cerebro humano, para analizar y modelar grandes volúmenes de datos complejos.

Redes Neuronales: Son estructuras inspiradas en el cerebro humano que permiten a las máquinas "aprender" de los datos, ajustando sus conexiones internas en función de la experiencia.

LLMs (Modelos de Lenguaje de Gran Escala): Son sistemas de inteligencia artificial entrenados con enormes cantidades de texto, capaces de comprender y generar lenguaje humano de manera fluida, como es el caso de ChatGPT.

Agente de Inteligencia Artificial: Es un sistema que combina modelos de IA con herramientas que permiten ejecutar tareas más complejas, como buscar información, tomar decisiones o interactuar con los usuarios.

Prompt: Es la instrucción o pregunta que se le proporciona a un modelo de lenguaje para que genere una respuesta.

Transformers: Es una arquitectura de inteligencia artificial altamente eficiente para el procesamiento de texto y lenguaje, que constituye la base de los modelos de lenguaje modernos como ChatGPT.

RAG (Retrieval-Augmented Generation): Es una técnica que combina modelos de inteligencia artificial con sistemas de búsqueda en bases de datos o textos, para ofrecer respuestas más precisas y actualizadas.

Fine-tunning: Es el proceso de ajuste fino que se realiza sobre un modelo preexistente, utilizando datos específicos (por ejemplo, textos médicos), para adaptarlo mejor a una tarea concreta.

Alucinaciones: Son errores de los modelos de inteligencia artificial cuando generan información incorrecta, pero la presentan como si fuera verdadera.

Embeddings: Son representaciones numéricas de palabras, frases o conceptos, que permiten a la inteligencia artificial entender similitudes y relaciones entre ellos.

Inferencia: Es el proceso mediante el cual un modelo de inteligencia artificial utiliza lo que ha aprendido para generar una respuesta, hacer una predicción o resolver una tarea nueva.

Multimodalidad: Es la capacidad de un modelo de inteligencia artificial para procesar distintos tipos de datos de manera simultánea, como texto, imágenes, audio o video.

Introducción

La inteligencia artificial (IA) representa una de las transformaciones más significativas en el ámbito de la salud digital y la cardiología contemporánea. Durante los últimos años, hemos sido testigos de un avance acelerado en el desarrollo y capacidades de estos sistemas, que han pasado de herramientas experimentales a soluciones con aplicaciones clínicas tangibles. Los modelos han evolucionado exponencialmente, logrando que arquitecturas generalistas y multidominio puedan realizar tareas complejas sin entrenamiento específico, incluso en áreas médicas altamente especializadas. Este documento aborda de manera estructurada las consideraciones fundamentales sobre la IA en el contexto sanitario, organizándose inicialmente en recomendaciones para el uso personal de modelos de IA generativa, seguido por lineamientos para su implementación institucional. Adicionalmente, se presenta un análisis exhaustivo de las métricas de evaluación de estos sistemas, para concluir con consideraciones éticas y legales que todo profesional de la salud debe contemplar. Cabe destacar que los aspectos específicos de la IA aplicada a imágenes cardiovasculares serán tratados en una sección independiente dentro de este consenso, dada su relevancia y particularidades técnicas.

Recomendaciones para el uso personal de modelos LLMs y de la IA generativa

Los Modelos de Inteligencia Artificial avanzados han evolucionado desde sistemas centrados únicamente en el análisis de texto hacia modelos multimodales capaces de interpretar y generar contenido en diversos formatos (texto, imágenes, audio y video). Estos modelos, conocidos como Modelos de Lenguaje Extenso (LLMs) están demostrando un potencial significativo en diversos campos de la medicina, incluyendo la cardiología (1,2). Es destacable que algunos de estos modelos han demostrado superar evaluaciones médicas estandarizadas de diferentes países con calificaciones elevadas, evidenciando un sólido conocimiento médico que continúa mejorando constantemente con cada nueva interacción (3,4).

En la práctica clínica, estos modelos tienen potencial para optimizar el flujo de trabajo, mejorar la seguridad del paciente y reducir errores médicos (5). Estudios recientes sugieren que podrían mejorar la estructuración y síntesis de información médica (6). Actualmente se está evaluando la escritura de notas médicas a partir de la transcripción automática de audio durante la consulta, aunque su precisión y utilidad clínica real aún requieren validación rigurosa (7). La generación automatizada de resúmenes de historias clínicas también está en etapas iniciales de investigación (8).

Estos modelos multimodales pueden facilitar la generación de ideas en diversos contextos creativos, promoviendo la innovación no solo en la educación médica sino también en la formulación de nuevas hipótesis de investigación, protocolos clínicos y enfoques terapéuticos (9-11). Han demostrado ser de utilidad en la planificación de conferencias científicas, diseño de material educativo y generación de guías interactivas para estudiantes de medicina (12). Las herramientas basadas en IA multimodal pueden producir gráficos detallados e interpretar imágenes médicas sobre anatomía, fisiología y estudios complementarios para asistir a los profesionales de la salud y facilitar la educación de los pacientes (13,14).

Estos sistemas también han mejorado la comunicación con el paciente, facilitando su comprensión y promoviendo la adherencia al tratamiento mediante recursos visuales y auditivos adaptados a sus necesidades (15,16). Un aspecto destacable es la capacidad de estos modelos para proporcionar respuestas empáticas y adecuadamente formuladas a preguntas de pacientes, lo que podría complementar la comunicación médico-paciente, especialmente en situaciones donde los profesionales disponen de tiempo limitado para la consulta (17).

Sin embargo, los modelos multimodales no están exentos de sesgos y fallos en su diseño, por lo que no deben utilizarse sin supervisión médica rigurosa. Ante la falta de información

Recomendaciones sobre el uso de LLM e IA generativa	Clase	Nivel de evidencia
Corrección y Esquematización: se acepta utilizar modelos grandes de lenguaje (LLMs) en tareas de corrección de estilo, ortografía, traducciones, resúmenes y esquematización de información. (1,2)	I	C
Ejercicios Creativos: se acepta utilizar modelos de IA para ejercicios creativos como "lluvias de ideas" y juegos de roles. (3,4)	I	C
Generación de Imágenes: se acepta utilizar modelos de IA generativa para la creación de imágenes con fines ilustrativos. (5,6)	I	C
Búsqueda de Datos Fácticos: se desaconseja el uso de LLMs para la búsqueda de datos fácticos o en tareas donde no se pueda supervisar la información generada por los modelos. (7,8)	III	C
Toma de Decisiones Autónomas: no utilizar modelos de lenguaje de manera autónoma para la toma de decisiones sin supervisión humana, tanto en el diagnóstico como en el triaje médico o intervenciones terapéuticas. (9,10)	III	C
Supervisión de Información Generada: se desaconseja utilizar información generada por LLMs y otros modelos de IA sin supervisión previa. (11,12)	III	C
Protección de Datos Sensibles: se desaconseja incluir datos sensibles de pacientes en los modelos de IA o sus instrucciones (prompts) (13,14)	III	C

o entrenamiento inadecuado, estos pueden generar "alucinaciones" (resultados incorrectos o engañosos).

Se desaconseja enfáticamente el uso de estos sistemas para la recuperación de datos fácticos en salud debido a su tendencia a generar respuestas inexactas o sesgadas. Los resúmenes y notas médicas generados automáticamente no siempre son confiables y requieren verificación. A pesar de su capacidad para procesar grandes volúmenes de información, estos modelos (en su versión base y sin herramientas específicas) no están diseñados para verificar la veracidad de los datos ni para acceder a fuentes actualizadas de evidencia clínica. Su capacidad de autoreconocer si poseen información confiable en sus datos de entrenamiento es limitada. La falta de interpretabilidad y su propensión a errores justifican que cualquier implementación como sistema de apoyo en la toma de decisiones deba realizarse bajo estrictos protocolos de validación y auditoría para garantizar su seguridad y eficacia (18). La supervisión y verificación cruzada con fuentes primarias sigue siendo fundamental en entornos clínicos.

BIBLIOGRAFÍA

1. Meng X, Yan X, Zhang K, Liu D, Cui X, Yang Y, et al. The application of large language models in medicine: A scoping review. iScience 2024;27:109713. https://doi.org/10.1016/j.isci.2024.109713
2. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. NPJ DigitMed 2022;5:194. https://doi.org/10.1038/s41746-022-00742-2
3. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assistedmedical education using large language models. PLOS Digit Health 2023;2:e0000198. https://doi.org/10.1371/journal.pdig.0000198
4. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620:172-80.https://doi.org/10.1038/s41586-023-06291-2
5. Doreswamy N, Horstmanshof L. Generative AI Decision-Making Attributes in Complex Health Services: A Rapid Review. Cureus2025;17:e78257. https://doi.org/10.7759/cureus.78257
6. Chen Q, Hu Y, Peng X, Xie Q, Jin Q, Gilson A, et al Benchmarking large language models for biomedical natural language processing applications and recommendations. Nat Commun 2025;16:3280. https://doi.org/10.1038/s41467-025-56989-2
7. Smith M, Liang AS, Delahaie C, Hsia C, Shanafelt T, Pfeffer MA, et al. Ambient artificial intelligence scribes: physician burnout and perspectives on usability and documentation burden. J Am Med Inform Assoc 2025;32:375-80. https://doi.org/10.1093/jamia/ocae295.
8. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform 2022;23:bbac409. https://doi.org/10.1093/bib/bbac409
9. Ruksakulpiwat S, Kumar A, Ajibade A. Using ChatGPT in Medical Research: Current Status and Future Directions. J Multidiscip Healthc 2023;16:1513-20. https://doi.org/10.2147/JMD H.S413470
10. Peng C, Yang X, Chen A, Smith KE, PourNejatian N, Costa AB, et al. A study of generative large language model for medical research and healthcare. NPJ Digit Med 2023;6:210. https://doi.org/10.1038/s41746-023-00958-w
11. Lower K, Seth I, Lim B, Seth N. ChatGPT-4: Transforming Medical Education and Addressing Clinical Exposure Challenges in the Postpandemic Era. Indian J Orthop 2023;57:1527-44. https://doi.org/10.1007/s43465-023-00967-7
12. Moore S, Tong R, Singh A, Liu Z, Hu X, Lu Y, et al. Empowering Education with LLMs-The Next-Gen Interface and Content Generation. Springer 2023:32-37. https://doi.org/10.1007/978-3-031-36336-8_4
13. Hanycz SA, Antiperovitch P. A practical review of generative AI in cardiac electrophysiology medical education. J Electrocardiol 2025;90:153903. https://doi.org/10.1016/j.jelectrocard.2025.153903
14. Ganjavi C, Melamed S, Biedermann B, Eppler MB, Rodler S, Layne E, et al. Generative artificial intelligence in oncology. Curr Opin Urol 2025;35:205-13. https://doi.org/10.1097/MOU.0000000000001272
15. Carl N, Haggenmüller S, Wies C, Nguyen L, Winterstein JT, Hetz MJ, et al. Evaluating interactions of patients with large language models for medical

information. BJU Int 2025;135:1010-7. https://doi.org/10.1111/bju.16676

16. Singh S, Errampalli E, Errampalli N, Miran MS. Enhancing Patient Education on Cardiovascular Rehabilitation with Large Language Models. Mo Med 2025;122:67-71.
17. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183:589-96. https://doi.org/10.1001/jamainternmed.2023.1838.
18. Minssen T, Vayena E, Cohen IG. The Challenges for Regulating Medical Use of ChatGPT and Other Large Language Models. JAMA 2023;330:315-6. A. 2023;330:974. https://doi.org/10.1001/jama.2023.16286

Recomendaciones para Instituciones de Salud sobre Implementación y Desarrollo de Inteligencia Artificial en Cardiología

La incorporación temprana de Inteligencia Artificial (IA) en las instituciones constituye un elemento fundamental para anticiparse a los desafíos tecnológicos actuales, facilitando una transformación digital efectiva en el ámbito de la salud cardiovascular (1,2). Diversos estudios recientes demuestran que los principales obstáculos para una implementación exitosa no radican principalmente en aspectos técnicos, sino en factores culturales y organizacionales (2,3). Experiencias previas sugieren que las instituciones cardiológicas que fomentan una cultura receptiva a la innovación, con liderazgo participativo y comunicación fluida entre equipos, tienden a facilitar la integración de estas tecnologías de manera más armoniosa (4).

Las instituciones que logran integrar tempranamente la IA en su práctica cardiológica tienen mayores posibilidades de maximizar beneficios clínicos específicos, optimizar recursos, mejorar la calidad asistencial en pacientes cardíacos y alcanzar una ventaja competitiva significativa frente a aquellas que retrasan esta incorporación tecnológica (4). Asimismo, la creación de estructuras organizacionales específicas, como comités interdisciplinarios que involucren a cardiólogos, profesionales informáticos, especialistas en ética y pacientes con enfermedades cardiovasculares para evaluar el impacto clínico, la usabilidad y la aceptación de estas tecnologías, resulta fundamental para gestionar adecuadamente el desarrollo e implementación de la IA en cardiología (5). Finalmente, adoptar estrategias pragmáticas como la realización de proyectos piloto iniciales específicos en cardiología, así como la adaptación de modelos preexistentes al contexto clínico cardiovascular local, permite mitigar riesgos y facilita una implementación exitosa (5,6).

Recomendaciones para instituciones sobre la implementación y desarrollo de IA en cardiología	Clase	Nivel de evidencia
Crear un comité interdisciplinario específico en cardiología que supervise responsablemente la implementación y uso de IA. (5)	I	C
Realizar proyectos piloto específicos en cardiología antes del despliegue completo para evaluar eficacia, seguridad y aceptación. (6)	I	C
Se recomienda validar y ajustar los modelos de IA al contexto cardiológico local y al tipo de datos cardiológicos antes de su uso clínico. (5)	I	C
Utilizar preferentemente modelos pre-entrenados adaptados específicamente al contexto cardiológico antes de desarrollar modelos desde cero, optimizando recursos y tiempo. (7)	I	C
No utilizar IA autónoma sin supervisión humana para decisiones críticas en cardiología, como triage de emergencias cardiovasculares o intervenciones invasivas cardíacas. (8)	III	C
Informar claramente a pacientes cardiológicos y usuarios cuando interactúan con sistemas de IA aplicados a su cuidado cardiovascular. (9)	I	C
Priorizar el uso de modelos especializados en medicina cardiovascular, como Med-Gemini, Bio-Mistral o generalistas probados clínicamente (GPT-4) (7-10)	I	C

BIBLIOGRAFÍA

1. Hassan M, Kushniruk A, Borycki E. Barriers to and Facilitators of Artificial Intelligence Adoption in Health Care: Scoping Review. JMIR Human Factors. 2024;11:e48633. https://doi.org/10.2196/48633.
2. Krumholz HM. Guiding the AI Revolution in Cardiovascular Medicine. J Am Coll Cardiol. 2025;86:843-5. https://doi.org/10.1016/j.jacc.2025.08.021.
3. Kamel Rahimi A, Pienaar O, Ghadimi M, Canfell OJ, Pole JD, Shrapnel S, et al. Implementing AI in Hospitals to Achieve a Learning Health System: Systematic Review. J Med Internet Res. 2024;26:e49655. https://doi.org/10.2196/49655.
4. Loriga B, Baglivo F, Bellini V, Adembri C, Montomoli J, Cascella M, et al. Top three priorities for artificial intelligence integration into emergency, critical, and perioperative medicine: an interdisciplinary clinical expert consensus. J Anesth Analg Crit Care. 2025;5:77.
5. Lekadir K, Feragen A, Fofanah AJ, Glocker B, Cintas C, Langlotz CP, et al.; FUTURE-AI Consortium. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. 2025;388:e081554.
6. Lovejoy CA, Arora A, Buch V, Dayan I. Key considerations for the use of artificial intelligence in healthcare and clinical research. Future Healthc J 2022 ;9:75-8. https://doi.org/10.7861/fhj.2021-0128
7. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to Diagnose Complex Clinical Cases. NEJM AI. 2023;1(1). https://doi.org/10.1056/AIp2300031
8. World Health Organization (WHO). Ethics and governance of artificial intelligence for health. WHO; 2021. Available at: https://www.who. int/publications/i/item/9789240029200
9. UNESCO. Recomendación sobre la ética de la inteligencia artificial. UNESCO; 2021. Disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000377897
10. Pierri MD, Galeazzi M, D'Alessio S, Dottori M, Capodaglio I, Corinaldesi C, Marini M, Di Eusanio M. Evaluating Large Language Models in Cardiology: A Comparative Study of ChatGPT, Claude, and Gemini. Hearts. 2025;6(3):19. https://doi.org/10.3390/hearts6030019

Evaluación de modelos de inteligencia artificial y machine learning

La evaluación de modelos de IA en medicina es un proceso crítico para determinar si un modelo es confiable, preciso y útil en la práctica clínica. El campo de la IA en cardiología es muy amplio y, antes de evaluar un modelo, es importante entender qué tipo de modelo es y qué tarea realiza, ya que en base a ello seleccionaremos las métricas para cuantificar el desempeño.

Inicialmente, podemos clasificar a los modelos por el tipo de tarea que realizan, sin perder de vista que para cada tarea existen distintas arquitecturas o tipos de algoritmos que pueden ser utilizados. Así, en cardiología, los modelos de machine learning e inteligencia artificial comúnmente utilizados se pueden clasificar de la siguiente manera (1,2).

- Modelos de Clasificación: su tarea es brindar un resultado dentro de un número acotado de opciones. Por ejemplo, decir si un paciente tiene o no una enfermedad cardíaca (clasificación binaria) o determinar qué tipo de arritmia tiene un paciente en base al electrocardiograma (clasificación multiclase).
- Modelos de Regresión: su tarea es brindar un resultado en forma de número (valores continuos). Por ejemplo, predecir el volumen regurgitante en insuficiencia mitral en base a imágenes de ecocardiografía.
- Modelos de Detección de Anomalías: su tarea es detectar valores o características que están fuera de lo normal. Sirven tanto para identificar patrones anormales en estudios diagnósticos como ECG, como para detectar errores en prescripción de medicamentos o dosificaciones.
- Modelos de Segmentación: su tarea en cardiología es delimitar regiones anatómicas en estudios por imágenes.
- Son la base de los modelos de cuantificación automática. Permiten, por ejemplo,

delimitar áreas de interés en imágenes de resonancia magnética cardíaca (ventrículos, miocardio, aurículas, etc.) para realizar cuantificación automática de volúmenes, masa y función ventricular.

- Modelos de Procesamiento del Lenguaje Natural: comúnmente abreviados como LLMs (large language models), se caracterizan por su capacidad para entender, interpretar y generar texto en lenguaje humano. Estos modelos son utilizados en cardiología para extraer información útil de historias clínicas electrónicas, generar reportes médicos automatizados, asistir en la toma de decisiones clínicas y facilitar la comunicación médico-paciente, contribuyendo así a mejorar la eficiencia y calidad en la atención médica.

Para cada tipo de modelo existen herramientas cuantitativas (denominadas métricas) que permiten medir el desempeño objetivamente. A continuación, describimos las métricas más utilizadas en cada tipo de modelo y algunos parámetros para su interpretación.

Métricas para Modelos de Clasificación

En modelos para clasificación binaria (dos clases posibles: positivo/negativo, enfermo/sano), las métricas más comunes (1) para evaluar problemas de clasificación son:

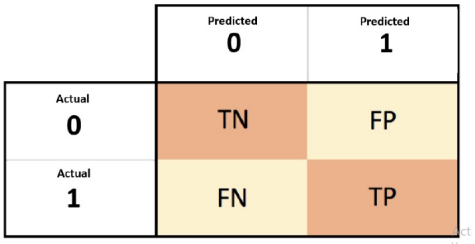
- Exactitud (Accuracy):
- Fórmula:

$$\text{Exactitud} = \frac{\text{Número de predicciones correctas}}{\text{Número total de predicciones}}$$

- Interpretación: Proporción de predicciones correctas sobre el total de casos. Es útil cuando las clases están balanceadas.
- Limitaciones: En cardiología, muchas enfermedades son poco prevalentes (clases desbalanceadas), lo que puede hacer que la exactitud sea engañosa. Por ejemplo, si solo el 3% de los pacientes tiene una enfermedad, un modelo que siempre predice "negativo" tendría una exactitud/accuracy del 97%, pero sería clínicamente inútil. Siempre que se utilice esta métrica para justificar la utilidad de un modelo se debería revisar cómo se trabajó en el balanceo de clases ya que esto puede limitar su utilidad o aplicación para otras poblaciones (3).

Matriz de confusión:

- La matriz de confusión es una herramienta visual que muestra cómo se distribuyen las predicciones del modelo en comparación con las etiquetas reales.
- Estructura: Las filas representan las clases reales, las columnas representan las clases predichas, la diagonal principal muestra los verdaderos positivos (VP) para cada clase.
- Interpretación: Permite identificar qué clases se confunden con mayor frecuencia. Es útil para detectar sesgos o problemas específicos en el modelo.



Sensibilidad (Recall o Tasa de Verdaderos Positivos):

• Fórmula:

$$\text{Sensibilidad} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

• Interpretación: Proporción de casos positivos correctamente identificados.

• Valor deseado: Lo más cercano a 1 posible.

Especificidad:

• Fórmula:

$$\text{Especificidad} = \frac{\text{VN}}{\text{VN} + \text{FP}}$$

• Interpretación: Proporción de casos negativos correctamente identificados.

• Valor deseado: Lo más cercano a 1 posible.

- Precisión (Precision):

• Fórmula:

$$\text{Precisión} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

• Interpretación: Proporción de predicciones positivas que son correctas. Es útil cuando los falsos positivos son costosos (por ejemplo, tratamientos innecesarios).

• Valor deseado: Lo más cercano a 1 posible.

- F1-Score:

• Fórmula:

$$\text{F1} = 2 \times \frac{\text{Precisión} \times \text{Sensibilidad}}{\text{Precisión} + \text{Sensibilidad}}$$

• Interpretación: Media armónica entre precisión y sensibilidad. Es útil cuando hay un desbalance entre clases.

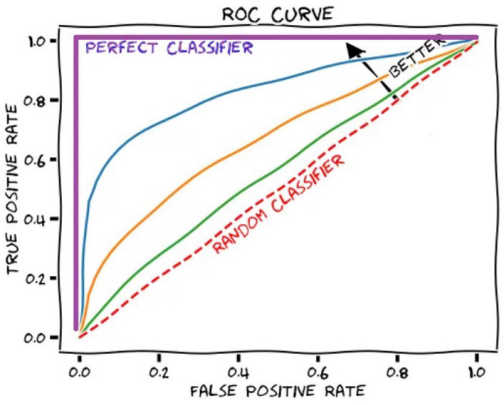
• Valor deseado: Lo más cercano a 1 posible.

- Curva ROC y Área bajo la Curva (AUC-ROC):

• Interpretación: La curva ROC grafica la sensibilidad frente a (1 - especificidad) para diferentes umbrales de clasificación. El AUC-ROC mide la capacidad del modelo para distinguir entre clases.

• Valor deseado: Un AUC-ROC de 0.5 indica un modelo no mejor que el azar, mientras que un valor de 1 indica un modelo perfecto.

En modelos de clasificación multiclase, métricas como exactitud, precisión, sensibilidad, especificidad y F1-score se calculan individualmente por clase y pueden promediarse (macro o ponderado). Otras medidas útiles incluyen el AUC-ROC multiclase (OvR u OvO) y el coeficiente Kappa, que evalúa concordancia ajustada por azar, especialmente útil con clases desbalanceadas.



Métricas para Modelos de Regresión

En problemas de regresión, cuyo resultado consiste en un número (valores continuos), las métricas más comunes (4) para evaluar la performance de un modelo son:

Error Cuadrático Medio (MSE):

• Fórmula:

• Interpretación: Mide el promedio de los errores al cuadrado entre los valores reales y los predichos.

• Valor deseado: Lo más cercano a 0 posible.

• No se suele utilizar en la práctica ya que al elevar los errores al cuadrado, el error no tiene la misma magnitud ni unidad que la variable predicha.

- Raíz del Error Cuadrático Medio (RMSE):

• Fórmula:

• Interpretación: Tanto MSE como RMSE son útiles cuando los errores grandes son especialmente indeseables.

RMSE da más peso a los errores grandes debido al cuadrado de las diferencias. Está en las mismas unidades que la unidad predicha, por lo cual su interpretación es más sencilla y suele ser el score más utilizado.

• Valor deseado: Lo más cercano a 0 posible.

• Ejemplo: si el MSE fue 100, la raíz sería 10. Esto significa que, en promedio, te equivocaste por 10 unidades al medir el nivel de glucosa en esta evaluación del modelo.

Tanto MSE como RMSE son útiles cuando los errores grandes son especialmente indeseables. RMSE da más peso a los errores grandes debido al cuadrado de las diferencias.

Ejemplo: Si el MSE fue 100, la raíz sería 10. Esto significa que, en promedio, te equivocaste por 10 unidades al medir el nivel de glucosa en esta evaluación del modelo.

- Error Absoluto Medio (MAE):

• Fórmula:

• Interpretación: Mide el promedio de los errores absolutos.

• Valor deseado: Lo más cercano a 0 posible.

• Es más robusto a los valores atípicos y ofrece una medida más directa de la cantidad "promedio" en que las predicciones se desvían de los valores reales (5).

- Coeficiente de Determinación (R^2):

• Fórmula:

• Interpretación: Proporción de la varianza en la variable dependiente que es predecible a partir de las variables independientes. Proporciona una medida de cuán bien las predicciones del modelo se aproximan a la variabilidad de los datos observados en el modelo entrenado.

• Valor deseado: Lo más cercano a 1 posible. Un valor de R^2 cercano a 1 indica que el modelo explica una gran proporción de la variabilidad en la variable objetivo1.

Métricas para Modelos de Segmentación

En problemas de segmentación la evaluación puede ser realizada con métricas denominadas geométricas, que buscan comparar la similitud o diferencia entre la imagen que debiera haber predicho el modelo y la que realmente generó. Las métricas de similitud geométrica más comunes (6) son:

- Coeficiente Dice (F1-Score para segmentación):

• Fórmula:

• Interpretación: Mide la superposición entre la segmentación predicha y la real.

• Valor deseado: Lo más cercano a 1 posible.

- Índice de Jaccard:

• Fórmula:

• Interpretación: Similar al Coeficiente Dice, pero menos sensible a superposiciones pequeñas.

• Valor deseado: Lo más cercano a 1 posible.

- Distancia de Hausdorff

• Interpretación: mide la mayor de las distancias mínimas entre puntos de dos conjuntos de datos (en este caso la segmentación predicha y la real). Es una métrica enfocada en la diferencia de los bordes y no en la superposición (a diferencia del coeficiente Dice y el índice de Jaccard).

• Valor deseado: Lo más cercano a 0 posible.

Además de la similitud geométrica, los modelos de segmentación pueden ser evaluados en base a los resultados de las cuantificaciones derivadas de sus predicciones. Por ejemplo, un modelo para medir volumen del ventrículo izquierdo puede ser evaluado con métricas como: diferencia de medias, correlación, concordancia y correlación intraclase.

Evaluación de modelos de lenguaje de procesamiento del lenguaje natural

Para evaluar el rendimiento general de un modelo de LLM se utilizan benchmarks amplios, es decir, pruebas estandarizadas que cubren múltiples temas. Por ejemplo, MMLU (Massive Multitask Language Understanding) es un benchmark con preguntas de opción múltiple en 57 materias diferentes, desde ciencias y matemáticas hasta humanidades, abarcando niveles desde primaria hasta nivel profesional (7). Otro ejemplo es GSM8K (Grade School Math 8K), un conjunto de 8.500 problemas de matemáticas de escuela primaria diseñados para medir la capacidad de razonamiento aritmético paso a paso de los modelos (8). Estos benchmarks generales sirven para medir las habilidades amplias del LLM (conocimiento general y lógica), pero no necesariamente reflejan su competencia en un dominio específico como la medicina.

En el ámbito de la salud, se emplean benchmarks específicos que contienen preguntas y tareas directamente relacionadas con temas médi-

cos. En este caso el benchmark más robusto es el **MultiMedQA** introducido por Google, que combina seis conjuntos de datos de preguntas médicas (incluyendo preguntas de exámenes médicos profesionales, de investigación biomédica y consultas de pacientes), además de un nuevo conjunto de preguntas médicas buscadas en la web (HealthSearchQA) (9). Este tipo de evaluación especializada, utilizada por ejemplo para el modelo Med-PaLM de Google, permite medir con mayor precisión cómo responde el LLM a cuestiones clínicas reales, como interpretar resultados, diagnosticar a partir de síntomas o contestar dudas de pacientes, ámbitos donde el lenguaje y el contexto son muy específicos.

La exactitud, medida por un número o porcentaje de preguntas contestadas correctamente por el modelo, es una métrica central en la evaluación cuantitativa de los LLMs. Al interpretarla, es esencial considerar dos conceptos fundamentales. En primer lugar, esta exactitud siempre debe ser contextualizada mediante la comparación con el desempeño humano, evaluando qué porcentaje de respuestas correctas obtiene una población representativa de profesionales o expertos en la misma tarea, dado que el rendimiento humano constituye un estándar crítico para interpretar los resultados del modelo. En segundo lugar, se debe considerar la variabilidad intrínseca en el rendimiento del modelo a través de múltiples pruebas o repeticiones, ya que las respuestas pueden variar según pequeños cambios en el contexto o la formulación de las preguntas. Esta variabilidad exige realizar varios tests paralelos para obtener una medición robusta y confiable. Finalmente, resulta útil y relevante comparar sistemáticamente la exactitud obtenida con versiones previas del mismo modelo o con modelos similares desarrollados por otras empresas o grupos de investigación, lo cual permite evaluar objetivamente la evolución del desempeño del LLM y posicionarlo claramente en el panorama actual del procesamiento del lenguaje natural en salud.

Más allá de la exactitud numérica en una prueba, en medicina interesa qué tan buena es la respuesta generada por el modelo en términos clínicos. Por eso, se han definido métricas cualitativas adicionales para evaluar LLMs médicos, las cuales suelen requerir la revisión por expertos humanos. Algunos aspectos clave son:

• Coherencia: que la respuesta esté bien estructurada, sea clara y tenga un hilo lógico fácil de seguir para el personal de salud y pacientes. Una respuesta coherente indica que el modelo entendió la pregunta y explica su razonamiento de forma ordenada.

• Precisión factual: que la información médica proporcionada sea correcta y verificable según el conocimiento científico vigente. Esto implica ausencia de errores médicos o afirmaciones sin fundamento; por ejemplo, que cite dosis, diagnósticos o datos epidemiológicos acertados y acordes al consenso médico (10).

• Utilidad clínica: que la respuesta aportada realmente sea útil en un contexto clínico, es decir, que ayude en la toma de decisiones médicas o en la educación del paciente. Una respuesta con alta utilidad clínica responde a la pregunta planteada de forma pertinente y accionable (por ejemplo, sugiriendo posibles pasos a seguir o consideraciones relevantes para el caso).

Además de las anteriores, a menudo se consideran otros criterios como la completitud (que no omita información importante), la ausencia de sesgos o contenido inapropiado y la alineación con prácticas seguras. Por ejemplo, en la evaluación de **Med-PaLM** se incluyeron ejes como la concordancia con el consenso científico, la ausencia de potencial daño en las recomendaciones y la utilidad percibida de la respuesta por profesionales de la salud humanos (9). Estas métricas específicas buscan reflejar la calidad real de las respuestas de un LLM en un escenario clínico, algo que porcentajes crudos de aciertos no capturan por sí solos.

Seguridad en modelos de procesamiento de lenguaje natural: Un aspecto crítico al usar LLMs en salud es garantizar su seguridad, es decir, que no generen contenido dañino o inapropiado aunque se les intente manipular. El término "jailbreak" se refiere precisamente a técnicas con las que un usuario malintencionado

puede "engañar" al modelo para que salte sus restricciones de seguridad y produzca respuestas que normalmente tendría prohibido dar (11). Por ejemplo, mediante indicaciones astutas, alguien podría inducir a un LLM a revelar información confidencial, a dar consejos médicos peligrosos o a usar un lenguaje inapropiado, cosas que el modelo no haría bajo instrucciones normales. En el contexto médico, un jailbreak es especialmente preocupante porque podría lograr que el modelo brinde información errónea o no ética (p. ej., recomendar un tratamiento contraindicado o sin evidencia) que el sistema debería haber rechazado).

Para reducir estos riesgos, los desarrolladores implementan múltiples medidas de seguridad. Primero, entrenan a los modelos con técnicas de refuerzo y filtros para que reconozcan peticiones inapropiadas y se nieguen a responder si el contenido es indebido (por ejemplo, si alguien pide instrucciones para un uso indebido de medicamentos, el modelo debería rechazarlo). También se recurre a pruebas adversariales (red teaming), donde se ponen a prueba deliberadamente los límites del LLM con intentos de jailbreak, detectando vulnerabilidades antes de su uso público. Otras estrategias incluyen filtros automáticos que monitorean las salidas del modelo en busca de lenguaje o consejos peligrosos. En general, la mitigación del jailbreak requiere un enfoque integral con vigilancia continua: sistemas de seguridad robustos, pruebas periódicas y actualizaciones del modelo para cerrar brechas (12). Esto ayuda a asegurar que, incluso ante usuarios maliciosos, el LLM mantenga comportamientos alineados con la ética y la seguridad clínica.

Alucinaciones en modelos de procesamiento del lenguaje natural: Otro desafío de los LLMs en general (no solo en el ámbito de la salud) son las alucinaciones, término que describe cuando el modelo genera con gran confianza información que no es correcta ni está basada en hechos reales. En la práctica, los LLMs pueden producir una respuesta que suena plausible y detallada, pero que en realidad contiene datos inventados o errores factuales (13). Por ejemplo, podría mencionar un estudio clínico que no existe, ofrecer una estadística incorrecta sobre una enfermedad, o recomendar un fármaco inexistente.

Este fenómeno ocurre porque el modelo no "sabe" en el sentido humano, sino que predice palabras según patrones aprendidos de sus datos de entrenamiento. Si en su entrenamiento hubo lagunas de información o contenidos ambiguos, el LLM tiende a rellenar esos vacíos con su mejor conjetura estadística, generando así información incorrecta sin darse cuenta. La ambigüedad en la pregunta o instrucciones del usuario también puede contribuir: si la consulta es vaga, el modelo puede divagar y salirse de los hechos.

En el campo de la salud, las alucinaciones son especialmente preocupantes. A diferencia de otras áreas donde una respuesta inexacta puede no tener mayores consecuencias, una afirmación médica incorrecta puede afectar decisiones clínicas o la seguridad del paciente (14). Por ejemplo, información falsa pero formulada de forma convincente podría llevar a un médico a considerar un diagnóstico equivocado o a un paciente a seguir un consejo inapropiado. Por este motivo, se están explorando varias estrategias para mitigar las alucinaciones en los LLMs. Una línea de trabajo es mejorar la formación del modelo con datos médicos de alta calidad y actualizados, e incluso ajustarlos para que cuando no estén seguros, expresen dudas en lugar de inventar respuestas. Otra estrategia es integrar al LLM con herramientas de búsqueda y verificación: por ejemplo, que el modelo pueda consultar bases de datos médicas o literatura científica en tiempo real para respaldar sus respuestas, en lugar de depender solo de su conocimiento interno. También se emplean técnicas de verificación factual automática y controles de consistencia: el sistema revisa lo que el LLM responde y lo compara con fuentes confiables o con otras formulaciones de la pregunta, detectando posibles contradicciones. En resumen, minimizar las alucinaciones requiere combinar mejores datos, ajustes en el comportamiento del modelo y apoyo de herramientas externas. Mientras tanto, siempre se recomienda la supervisión humana: un médico debe validar la información crucial

que ofrezca un LLM antes de aplicarla, usando su propio juicio clínico y fuentes médicas oficiales, de modo que las "alucinaciones" del modelo no se traduzcan en errores en la atención de los pacientes.

Consideraciones Éticas y Clínicas sobre Sesgos en Modelos de IA en Salud

Los modelos de inteligencia artificial en medicina pueden perpetuar desigualdades al entrenarse con datos sesgados que no representan adecuadamente a todas las poblaciones. Por ejemplo, modelos de riesgo cardiovascular, como los derivados del estudio Framingham (15) han mostrado menor precisión al estimar riesgos en pacientes afroamericanos debido a su desarrollo con datos principalmente caucásicos, generando diagnósticos tardíos o insuficientes. Similarmente, modelos que predicen insuficiencia cardíaca suelen fallar más en mujeres debido a su infrarrepresentación en estudios previos. Para mitigar estos sesgos, es fundamental asegurar la diversidad en los datos de entrenamiento y utilizar técnicas que limiten la influencia de variables sensibles como género o etnia. A nivel regulatorio, agencias como la FDA en Estados Unidos y organismos internacionales como la OMS han establecido recomendaciones para promover la transparencia, equidad y monitoreo constante del desempeño de estos modelos, garantizando así que beneficien equitativamente a todos los grupos poblacionales, evitando discriminación algorítmica en la práctica clínica (16).

Conclusión

La evaluación de modelos de IA en cardiología requiere un enfoque riguroso y contextualizado. Las métricas deben seleccionarse en función del tipo de modelo y del problema clínico, y su interpretación debe considerar las consecuencias de los errores. Además, la validación y las consideraciones éticas son fundamentales para garantizar que los modelos sean seguros y útiles en la práctica médica.

Recomendaciones sobre Evaluación de Modelos de AI:

• Evaluación de la Performance de Modelos de IA. Recomendación de Clase I, Nivel de evidencia C.

o La evaluación de la performance de un modelo de IA debe realizarse a partir de los resultados del dataset de testeo, con datos que el modelo nunca vio durante el entrenamiento. Esto garantiza una medida real de la performance, independiente de los resultados obtenidos en los datos de entrenamiento y validación.

• Métricas para Problemas de Clasificación. Recomendación de Clase I, Nivel de evidencia C.

o Para problemas de clasificación, se deben utilizar métricas de accuracy, precision, recall, F1 y matriz de confusión:

o Accuracy: Porcentaje de predicciones correctas.

o Precisión: Proporción de verdaderos positivos entre los positivos predichos.

o Recall: Proporción de verdaderos positivos entre los positivos reales.

o F1 Score: Media armónica de precisión y recall.

o Matriz de Confusión: Representación visual de las predicciones correctas e incorrectas.

• Métricas para Problemas de Regresión. Recomendación de Clase I, Nivel de evidencia C.

o Para problemas de regresión, se deben utilizar las métricas de MAE, MSE y RMSE:

o MAE (Mean Absolute Error): Promedio de los errores absolutos.

o MSE (Mean Squared Error): Promedio de los errores al cuadrado.

o RMSE (Root Mean Squared Error): Raíz cuadrada del MSE, da mayor peso a los grandes errores.

• Uso de Accuracy en Variables Desbalanceadas. Recomendación de Clase III, Nivel de evidencia C.

o Se desaconseja la utilización de accuracy como único parámetro de performance en casos

donde la variable a predecir esté desbalanceada, ya que puede generar una falsa sensación de éxito. En estos casos, es recomendable utilizar métricas adicionales como AUC-ROC, balance de precisión y recall, y el F1 Score.

• Evaluaciones en Modelos de LLMs. Recomendación de Clase I, Nivel de evidencia C.

o En modelos LLMs, las evaluaciones deben estar realizadas por humanos o, en su defecto, por LLMs que han demostrado tener evidencia comparable a la evaluación humana en ese tipo de tareas. Es importante validar las evaluaciones automáticas con estudios comparativos respecto a la evaluación humana para asegurar su precisión.

Recomendaciones sobre Evaluación de Modelos de AI	Clase	Nivel de evidencia
La evaluación de la performance de un modelo de IA debe realizarse a partir de los resultados del dataset de testeo, con datos que el modelo nunca vio durante el entrenamiento. Esto garantiza una medida real de la performance, independiente de los resultados obtenidos en los datos de entrenamiento y validación	I	C
Para problemas de clasificación, se deben utilizar métricas de accuracy, precision, recall, F1 y matriz de confusión	I	C
Para problemas de regresión, se deben utilizar las métricas de MAE, MSE y RMSE	I	C
Se desaconseja la utilización de accuracy como único parámetro de performance en casos donde la variable a predecir esté desbalanceada, ya que puede generar una falsa sensación de éxito. En estos casos, es recomendable utilizar métricas adicionales como AUC-ROC, balance de precisión y recall, y el F1 Score.	III	C
En modelos LLMs, las evaluaciones deben estar realizadas por humanos o, en su defecto, por LLMs que han demostrado tener evidencia comparable a la evaluación humana en ese tipo de tareas. Es importante validar las evaluaciones automáticas con estudios comparativos respecto a la evaluación humana para asegurar su precisión.	I	C

BIBLIOGRAFÍA

1. Peter Bruce, Andrew Bruce, Peter Gedeck - Practical Statistics for Data Scientists 50+ Essential Concepts Using R and Python-O'Reilly Media (2020)
2. Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions. BMJ. 2020;368:l6927. https://doi.org/10.1136/bmj.l6927.
3. Gurcan F, Soylu A. Learning from Imbalanced Data: Integration of Advanced Resampling Techniques and Machine Learning Models for Enhanced Cancer Diagnosis and Prognosis. Cancers (Basel) 2024;16):3417. https://doi.org/10.3390/cancers16193417
4. Aggarwal, C.C. (2024) Probability and Statistics for Machine Learning: A Textbook. Cham: Springer. https://doi.org/10.1007/978-3-031- 53282-5
5. Willmott CJ, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim Res 2005:30:79-82 https://doi.org/10.3354/cr030079
6. Eelbode T, Bertels J, Berman M, Vandermeulen D, Maes F, Bisschops R, et al. Optimization for Medical Image Segmentation: Theory and Practice When Evaluating With Dice Score or Jaccard Index. IEEE Trans Med Imaging 2020;39:3679-90. https://doi.org/10.1109/ TMI.2020.3002417
7. Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, et al. Measuring Massive Multitask Language Understanding. ArXiv 2020;arXiv:2009.03300.
8. Cobbe K, Kosaraju V, Bavarian M, Chen M, Jun H, Kaiser L, et al. Training Verifiers to Solve Math Word Problems. ArXiv 2021;abs/2110.14168.
9. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620:172-80. https://doi.org/10.1038/s41586-023-06291-2
10. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards Expert-Level Medical Question Answering with Large Language Models. ArXiv 2023;abs/2305.09617. https://doi.org/10.48550/arXiv.2305.09617.
11. Mondillo G, Colosimo S, Perrotta A, Frattolillo V, Indolfi C, Miraglia del Giudice M, et al. Jailbreaking large language models: navigating the crossroads of innovation, ethics, and health risks. Journal Of Medical Artificial Intelligence 202;8. https://doi.org/10.21037/jmai-24-170
12. Rao A, Vashistha S, Naik A, Aditya S, Choudhury M. Tricking LLMs into Disobedience: Understanding, Analyzing, and Preventing Jailbreaks. ArXiv 2023;abs/2305.14965. https://doi.org/10.48550/arXiv.2305.14965
13. Aditya G. Understanding and Addressing AI Hallucinations in Healthcare and Life Sciences. International Journal of Health Sciences 2024;7:1–11. https://doi.org/10.47941/ijhs.1862
14. Kim Y, Jeong H, Chen S, Stella Li S, Lu M, et al. Medical Hallucination in Foundation Models and Their Impact on Healthcare. medRxiv 2025.02.28.25323115; https://doi.org/10.1101/2025.02.28.2532311
15. Celi LA, Cellini J, Charpignon ML, Dee EC, Dernoncourt F, Eber R, et al; for MIT Critical Data. Sources of bias in artificial intelligence that perpetuate healthcare disparities-A global review. PLOS Digit Health 2022;1:e0000022. https://doi.org/10.1371/journal.pdig.0000022
16. Jain LR, Menon V. AI Algorithmic Bias: Understanding its Causes, Ethical and Social Implications. ICTAI 2023:460-7. https://doi. org/10.1109/ICTAI59109.2023.00073