noemicajournal.org

e-ISSN: 3086-7037

ARTÍCULO

Gobernanza responsable de la inteligencia artificial en sistemas de salud: principios para la equidad, la transparencia y la capacidad institucional

Governança responsável da inteligência artificial em sistemas de saúde: princípios para equidade, transparência e capacidade institucional

Responsible governance of artificial intelligence in healthcare systems: principles for equity, transparency, and institutional capacity

Antonio Pires Barbosa[1]

[1] Universidade Nove de Julho (UNINOVE), São Paulo, SP, Brasil. ORCID: https://orcid.org/0000-0001-6478-6522.

ENVIADO	ACEPTADO	PUBLICADO	DOI
12/03/2025	20/10/2025	05/12/2025	en proceso

RESUMEN

La incorporación de la inteligencia artificial a los sistemas de salud plantea desafíos de gobernanza que van más allá de la validación técnica de los algoritmos. Este artículo examina cómo los principios normativos pueden orientar la adopción responsable de sistemas de diagnóstico, triaje y gestión clínica, prestando especial atención a la equidad en la atención, la transparencia de las decisiones automatizadas y el fortalecimiento de la capacidad institucional dentro de las organizaciones de salud. La discusión se basa en una revisión integradora de la literatura sobre gobernanza clínica, seguridad del paciente, justicia distributiva, explicabilidad algorítmica y capacidades organizacionales, incorporando contribuciones publicadas en revistas como Nature Medicine, The New England Journal of Medicine, Science y BMC Medicine, así como las directrices de la Organización Mundial de la Salud sobre ética y gobernanza de la

inteligencia artificial para la salud. A partir de esta base, el artículo propone un modelo compuesto por seis dimensiones interrelacionadas: propósito clínico legítimo, calidad y representatividad de los datos, validación contextual, supervisión humana efectiva, mecanismos de impugnación y monitoreo continuo. Se argumenta que el desempeño técnico de un algoritmo, medido de forma aislada mediante métricas como la sensibilidad o el área bajo la curva ROC, no garantiza el beneficio público cuando los sistemas se implementan en redes de atención desiguales, con infraestructura, capacitación y recursos dispares. La contribución del artículo consiste en integrar requisitos técnicos, éticos, legales y administrativos en un marco aplicable a todo el ciclo de vida de la inteligencia artificial en la salud, desde la adquisición hasta su eventual retiro, ofreciendo orientación tanto para gestores de servicios como para formuladores de políticas.

Palabras clave: inteligencia artificial; sistemas de salud; gobernanza; equidad; transparencia; capacidad institucional.

RESUMO

A incorporação da inteligência artificial aos sistemas de saúde traz desafios de governança que vão além da validação técnica dos algoritmos. Este artigo examina como princípios normativos podem orientar a adoção responsável de sistemas de diagnóstico, triagem e gestão clínica, com atenção especial à equidade no atendimento, à transparência das decisões automatizadas e ao fortalecimento da capacidade institucional das organizações de saúde. A discussão baseia-se em uma revisão integrativa da literatura sobre governança clínica, segurança do paciente, justiça distributiva, explicabilidade algorítmica e capacidades organizacionais, incorporando contribuições publicadas em periódicos como Nature Medicine, The New England Journal of Medicine, Science e BMC Medicine, além das diretrizes da Organização Mundial da Saúde sobre ética e governança da inteligência artificial para a saúde. Com base nesse referencial, o artigo propõe um modelo composto por seis dimensões inter-relacionadas: propósito clínico legítimo, qualidade e representatividade dos dados, validação contextual, supervisão humana efetiva, mecanismos de contestação e monitoramento contínuo. Argumenta-se que o desempenho técnico de um algoritmo, medido isoladamente por métricas como sensibilidade ou área sob a curva ROC, não garante benefício público quando os sistemas são implementados em redes de atenção desiguais, com infraestrutura, treinamento e recursos díspares. A contribuição do artigo está em integrar exigências técnicas, éticas, legais

e administrativas em um arcabouço aplicável a todo o ciclo de vida da inteligência artificial na saúde, da aquisição ao eventual descomissionamento, oferecendo orientação tanto para gestores de serviços quanto para formuladores de políticas.

Palavras-chave: inteligência artificial; sistemas de saúde; governança; equidade; transparência; capacidade institucional.

ABSTRACT

The incorporation of artificial intelligence into healthcare systems raises governance challenges that go beyond the technical validation of algorithms. This article examines how normative principles can guide the responsible adoption of diagnostic, triage, and clinical management systems, with particular attention to equity in care, transparency of automated decisions, and the strengthening of institutional capacity within health organizations. The discussion draws on an integrative review of the literature on clinical governance, patient safety, distributive justice, algorithmic explainability, and organizational capabilities, engaging contributions published in journals such as Nature Medicine, The New England Journal of Medicine, Science, and BMC Medicine, as well as the World Health Organization guidance on ethics and governance of artificial intelligence for health. Building on this foundation, the article proposes a model composed of six interrelated dimensions: legitimate clinical purpose, data quality and representativeness, contextual validation, effective human oversight, contestation mechanisms, and continuous monitoring. It argues that the technical performance of an algorithm, measured in isolation through metrics such as sensitivity or the area under the ROC curve, does not guarantee public benefit when systems are deployed across unequal care networks with disparate infrastructure, training, and resources. The article's contribution lies in integrating technical, ethical, legal, and administrative requirements into a framework applicable across the entire lifecycle of artificial intelligence in healthcare, from procurement to eventual decommissioning, offering guidance for both service managers and policy makers.

Keywords: artificial intelligence; healthcare; governance; equity; transparency; institutional capacity.

INTRODUCCIÓN

Los sistemas de inteligencia artificial ya forman parte de la rutina de muchos servicios de salud. Apoyan la lectura de exámenes de imagen, ayudan a estratificar el riesgo cardiovascular, señalan a los pacientes con mayor probabilidad de deterioro clínico, organizan listas de espera y colaboran en la planificación de los recursos hospitalarios. Topol (2019) describió este movimiento como parte de una transición hacia una medicina de alto desempeño, en la que los algoritmos de aprendizaje automático procesan volúmenes de datos que superan la capacidad humana de análisis manual. Este cambio trae beneficios reales en ciertos contextos, pero también introduce riesgos que no se limitan a fallas técnicas del software.

El primer riesgo es el error clínico derivado de modelos mal calibrados o aplicados fuera del contexto para el cual fueron validados. Un algoritmo entrenado con datos de un gran hospital universitario puede comportarse de manera muy distinta cuando se utiliza en una unidad básica de salud que atiende a un perfil de pacientes diferente. Kelly et al. (2019), en una revisión publicada en BMC Medicine, mostraron que la mayoría de los estudios sobre inteligencia artificial en salud carecen de una validación externa robusta, lo que limita el grado en que los resultados reportados en los artículos técnicos pueden generalizarse a la práctica clínica cotidiana.

El segundo riesgo se refiere a la opacidad de los modelos. Las redes neuronales profundas y otros métodos de aprendizaje automático suelen funcionar como cajas negras, generando recomendaciones sin una explicación intuitiva del razonamiento subyacente. Esta característica dificulta la supervisión clínica y la rendición de cuentas cuando algo sale mal, un problema ya discutido por Verghese, Shah y Harrington (2018), quienes advirtieron que la confianza ciega en las recomendaciones computacionales puede debilitar, en lugar de fortalecer, el juicio médico.

El tercer riesgo, quizás el más estudiado en los últimos años, es la desigualdad. Obermeyer et al. (2019), en un artículo publicado en la revista Science, demostraron que un algoritmo ampliamente utilizado en Estados Unidos para identificar a los pacientes con necesidades de atención más complejas discriminaba sistemáticamente a los pacientes negros, porque usaba el gasto en salud como indicador indirecto de la necesidad clínica. Dado que los pacientes negros históricamente han tenido menos acceso a los servicios, incluso con condiciones de salud equivalentes, el modelo subestimaba la gravedad de sus cuadros. Este caso se ha convertido en una referencia obligada en cualquier discusión sobre la gobernanza de la inteligencia artificial en salud, porque muestra que un sistema con buenas

e-ISSN: 3086-7037

métricas de desempeño agregado puede, aun así, producir un daño distributivo grave.

Estos tres riesgos —el error clínico, la opacidad y la desigualdad— no son fallas de ingeniería que puedan resolverse simplemente con modelos más sofisticados. Se originan en decisiones institucionales sobre cómo se elige, implementa, supervisa y revisa la tecnología. Este es el argumento central de este artículo: la gobernanza de la inteligencia artificial en salud debe traducir principios éticos amplios en responsabilidades concretas, procedimientos verificables y criterios que permitan una auditoría posterior. Sin esta traducción, las decisiones relevantes tienden a trasladarse a los proveedores de tecnología o a especialistas aislados, lo que reduce la capacidad de la organización de salud para ejercer una supervisión efectiva sobre herramientas que afectan directamente la vida de los pacientes.

Un segundo hilo del argumento se refiere a la distribución desigual de capacidades entre las organizaciones de salud. Los hospitales universitarios, las grandes redes privadas y los sistemas públicos bien financiados cuentan con equipos de ciencia de datos, departamentos jurídicos y poder de negociación frente a los proveedores. Las unidades más pequeñas, especialmente en regiones con menor inversión en salud, generalmente carecen de esta estructura. Si la gobernanza de la inteligencia artificial se trata únicamente como un problema técnico que cada organización debe resolver internamente, el resultado probable es una concentración de beneficios en las instituciones que ya cuentan con más recursos, lo que ampliaría las desigualdades entre territorios y poblaciones. Las políticas de apoyo, los estándares compartidos y la infraestructura pública de evaluación pueden mitigar esta asimetría, pero requieren una decisión política deliberada.

Un tercer hilo se refiere a la necesidad de un monitoreo continuo. El desempeño de un sistema de inteligencia artificial no es estático. Los cambios en la población atendida, en los protocolos clínicos, en los equipos de captura de imágenes o incluso en los hábitos de codificación de las historias clínicas pueden alterar sustancialmente la calidad de las predicciones a lo largo del tiempo, un fenómeno conocido en la literatura técnica como deriva de datos (data drift). Sendak et al. (2020), en un estudio publicado en NPJ Digital Medicine sobre la implementación de un modelo predictivo de sepsis en un sistema hospitalario estadounidense, mostraron cómo el proceso de validación, ajuste y monitoreo continuo fue tan importante como el desarrollo inicial del algoritmo. Sin una revisión

periódica, un sistema validado en un momento dado puede volverse inadecuado sin que nadie lo note, hasta que el daño ya se haya producido.

Finalmente, vale la pena destacar la dimensión de la rendición de cuentas. Cuanto mayor sea el potencial de un sistema para afectar derechos, recursos u oportunidades de tratamiento, mayor debe ser el nivel de documentación, escrutinio y posibilidad de apelación disponible para los pacientes afectados. Esta proporcionalidad entre riesgo y control es un principio recurrente tanto en la literatura sobre regulación de dispositivos médicos como en las directrices recientes sobre inteligencia artificial, incluido el documento de la Organización Mundial de la Salud sobre ética y gobernanza de la inteligencia artificial para la salud (Organización Mundial de la Salud, 2021) y el enfoque basado en riesgo adoptado por el Reglamento Europeo de Inteligencia Artificial.

Este artículo se organiza de la siguiente manera. Tras esta introducción, se presenta el marco teórico que sustenta la propuesta, reuniendo la gobernanza clínica, la seguridad del paciente, la justicia distributiva y la teoría de las capacidades organizacionales. A continuación se desarrolla cada una de las seis dimensiones del modelo propuesto, discutiendo el propósito clínico, los datos, la validación, la transparencia, la supervisión humana, la impugnación y el monitoreo, así como la capacidad institucional. Una sección específica aborda la adquisición, la regulación, la capacitación profesional, la interoperabilidad y la evaluación económica, temas que a menudo quedan fuera de las discusiones estrictamente técnicas sobre inteligencia artificial en salud. El artículo cierra con una discusión ampliada sobre las implicaciones institucionales del modelo, un análisis de sus limitaciones y consideraciones finales sobre la agenda de investigación y de políticas.

MARCO TEÓRICO

La inteligencia artificial como tecnología sociotécnica

Un error común en las discusiones sobre inteligencia artificial en salud es tratar al algoritmo como un objeto aislado cuyo valor puede evaluarse únicamente mediante métricas estadísticas de desempeño. La literatura sobre sistemas sociotécnicos, aplicada a la informática en salud desde el trabajo clásico de Coiera (2007) sobre sistemas de información clínica, sostiene lo contrario: toda tecnología opera dentro de una red compuesta por personas, rutinas de trabajo, infraestructura física y reglas organizacionales. Un modelo de aprendizaje automático con un excelente

desempeño en condiciones de prueba puede fracasar por completo cuando se introduce en un flujo de trabajo que los profesionales no pueden seguir, o cuando genera alertas en un volumen que excede la capacidad de respuesta del equipo.

Cabitza, Rasoini y Gensini (2017), en un artículo publicado en JAMA sobre las consecuencias no deseadas del aprendizaje automático en medicina, llamaron la atención sobre el fenómeno de la complacencia con la automatización, en el cual los profesionales comienzan a aceptar las recomendaciones algorítmicas sin cuestionamiento crítico, incluso cuando el contexto clínico exige cautela. También ocurre lo contrario: los sistemas percibidos como poco confiables simplemente se ignoran, dejando inútil la inversión tecnológica. Ambos resultados se derivan de una mala integración sociotécnica, no necesariamente de una falla en el modelo en sí.

Esta perspectiva explica por qué la evaluación de un sistema de inteligencia artificial en salud no puede limitarse a pruebas de banco. Es necesario considerar cómo interpretan los profesionales las recomendaciones, cómo se inserta el sistema en las cargas de trabajo existentes y cómo responde la organización cuando el algoritmo se equivoca. Este es el trasfondo teórico sobre el cual se construyen las seis dimensiones propuestas más adelante.

Gobernanza clínica y seguridad del paciente

La tradición de la gobernanza clínica, consolidada en el Reino Unido a partir de la década de 1990 como respuesta a fallas graves en la atención, ofrece un vocabulario útil para pensar la inteligencia artificial en salud. La gobernanza clínica puede entenderse como el conjunto de estructuras mediante las cuales las organizaciones de salud rinden cuentas por la mejora continua de la calidad de sus servicios y por la salvaguarda de altos estándares de atención. Aplicada a la inteligencia artificial, esta tradición sugiere que la adopción de un algoritmo debe seguir el mismo rigor exigido para introducir un nuevo medicamento o procedimiento: evidencia de eficacia, evaluación de seguridad, capacitación del personal y monitoreo posterior a la implementación.

Wiens et al. (2019), en un artículo de referencia publicado en Nature Medicine titulado "Do no harm: a roadmap for responsible machine learning for health care", sistematizaron esta analogía al proponer una hoja de ruta para el desarrollo responsable del aprendizaje automático en salud, que abarca desde la formulación del problema clínico hasta el mantenimiento posterior a la implementación. Los autores sostienen que la mayoría de los problemas éticos asociados con la inteligencia artificial en salud no se

originan en la mala fe, sino en vacíos en los procesos, como la ausencia de validación prospectiva, la falta de monitoreo continuo y la escasez de mecanismos para reportar y corregir fallas.

Justicia distributiva y sesgo algorítmico

La discusión sobre la equidad en la inteligencia artificial en salud dialoga directamente con la tradición de la justicia distributiva en bioética, que se pregunta cómo deben distribuirse los beneficios y riesgos de una intervención entre los distintos grupos sociales. Los modelos de aprendizaje automático se entrenan con datos históricos, y los datos históricos reflejan las desigualdades de acceso, registro y calidad de la atención que ya existen dentro del sistema de salud. Cuando un grupo poblacional aparece con menor frecuencia, tiene registros de menor calidad o muestra resultados sistemáticamente peores en los datos de entrenamiento, el modelo tiende a reproducir, y en algunos casos a amplificar, esas diferencias.

El caso de Obermeyer et al. (2019), ya mencionado, ilustra claramente este mecanismo, pero está lejos de ser un caso aislado. Panch, Mattie y Atun (2019), en un artículo sobre inteligencia artificial y sesgo algorítmico publicado en el Journal of Global Health, sostienen que el sesgo puede entrar al ciclo de desarrollo de un sistema de inteligencia artificial en varias etapas, desde la elección del problema a resolver hasta la definición de las variables de resultado, pasando por la composición del conjunto de datos de entrenamiento y los criterios de validación. Por ello, los autores argumentan que la auditoría de equidad no puede ser un paso único añadido al final del desarrollo, sino que debe acompañar todo el proceso.

Rajkomar, Hardt, Howell, Corrado y Chin (2018), en un artículo publicado en Annals of Internal Medicine sobre la equidad en el aprendizaje automático aplicado a la salud, proponen una distinción útil entre diferentes nociones técnicas de equidad, como la paridad estadística, la igualdad de oportunidades y la calibración equilibrada entre subgrupos, mostrando que estas definiciones pueden entrar en conflicto entre sí y que la elección de un criterio de equidad es, en última instancia, un juicio de valor, no simplemente una cuestión técnica. Este hallazgo refuerza el argumento de este artículo de que la gobernanza de la inteligencia artificial en salud requiere una deliberación explícita sobre valores, y no solo la optimización de métricas.

Explicabilidad y confianza

La capacidad de explicar por qué un sistema llegó a una recomendación determinada suele tratarse como sinónimo de transparencia, pero la literatura reciente sugiere mayor cautela. Las explicaciones técnicas detalladas, como los mapas de saliencia en imágenes o los coeficientes de importancia de las variables, pueden resultar incomprensibles para profesionales sin formación en ciencia de datos, y aún más para pacientes legos. Amann et al. (2020), en un artículo sobre la explicabilidad de la inteligencia artificial en salud publicado en BMC Medical Informatics and Decision Making, sostienen que la explicabilidad debe entenderse en múltiples dimensiones, entre ellas la explicabilidad técnica, la justificación clínica y la comunicación adecuada a la audiencia, y que estas dimensiones responden a las necesidades diferentes de los distintos actores.

Desde el punto de vista de la confianza institucional, la transparencia cumple una función que va más allá de explicar cada predicción individual. Sustenta la posibilidad de una auditoría externa, la comparación entre sistemas competidores y la rendición de cuentas cuando ocurre un daño. Price y Cohen (2019), al discutir la privacidad y los macrodatos médicos en Nature Medicine, también advierten que la transparencia sobre el uso de los datos de los pacientes es una precondición para la legitimidad de cualquier sistema de inteligencia artificial en salud, especialmente cuando datos sensibles circulan entre instituciones públicas y privadas.

Capacidades organizacionales y teoría de la capacidad estatal

Finalmente, el modelo propuesto en este artículo se apoya en la literatura sobre capacidades organizacionales y capacidad estatal, que sostiene que la implementación efectiva de cualquier política o tecnología depende de recursos humanos calificados, infraestructura adecuada, procesos de coordinación y mecanismos de aprendizaje institucional. Sun y Medaglia (2019), en un estudio sobre la adopción de la inteligencia artificial en hospitales públicos chinos publicado en Government Information Quarterly, encontraron que las barreras organizacionales, como la resistencia profesional, la falta de estandarización de datos y la ausencia de un liderazgo comprometido, fueron tan relevantes como las limitaciones técnicas para explicar el éxito o el fracaso de los proyectos de inteligencia artificial en salud.

Esta literatura respalda el argumento de que la gobernanza responsable de la inteligencia artificial en salud no puede reducirse a una lista de verificación ética aplicada una sola vez. Requiere una capacidad continua de evaluación, ajuste y toma de decisiones, distribuida de manera desigual

entre las organizaciones, lo que convierte el apoyo de las políticas públicas a la capacidad institucional en un componente inseparable de la agenda ética y técnica.

PROPÓSITO CLÍNICO Y PROPORCIONALIDAD

El primer elemento del modelo propuesto es el requisito de un propósito clínico legítimo. Antes de discutir la arquitectura del modelo o la calidad de los datos, una organización necesita responder una pregunta sencilla: qué problema clínico u organizacional pretende resolver el sistema, para qué población y con qué beneficio esperado. Los sistemas adoptados sin esta definición clara tienden a evaluarse únicamente por su sofisticación técnica, la cual no corresponde necesariamente a su utilidad clínica.

La proporcionalidad se deriva directamente de este requisito. Un sistema de triaje que prioriza a los pacientes en la fila de un servicio de urgencias tiene un potencial de daño mucho mayor que una herramienta que organiza la agenda administrativa. Por ello, el nivel de evidencia exigido, la intensidad de la supervisión humana y el rigor de la documentación deben ser proporcionales al riesgo potencial de la aplicación, en lugar de uniformes para toda herramienta etiquetada como inteligencia artificial. Esta lógica de gradación basada en el riesgo está presente tanto en la regulación de los dispositivos médicos basados en software, como las directrices de la Food and Drug Administration para el software como dispositivo médico, como en el enfoque del Reglamento Europeo de Inteligencia Artificial, que clasifica a los sistemas utilizados en salud predominantemente como de alto riesgo, sujetos a requisitos reforzados de documentación, gestión de riesgos y supervisión humana.

Char, Shah y Magnus (2018), en un influyente artículo publicado en el New England Journal of Medicine sobre la implementación del aprendizaje automático en salud, advirtieron sobre el riesgo de adoptar tecnología por presión competitiva o por moda, sin que la organización tenga claridad sobre el problema que se pretende resolver. Los autores sostienen que las decisiones de adopción de la inteligencia artificial en salud deben seguir un proceso similar al utilizado para evaluar nuevas terapias, con etapas de recopilación de evidencia, ensayos controlados cuando sea posible y revisión por comités independientes antes de una adopción a gran escala.

En la práctica institucional, este requisito se traduce en documentos que registran el propósito, la población destinataria, las alternativas consideradas, el beneficio esperado y los criterios de éxito, aprobados antes

de adquirir el sistema. Este registro cumple una doble función: orienta la elección entre proveedores competidores y crea una línea de base para la evaluación posterior, permitiendo comparar lo que se prometió con lo que efectivamente ocurrió tras la implementación.

DATOS, REPRESENTATIVIDAD Y EQUIDAD

Los modelos de aprendizaje automático aprenden patrones a partir de datos históricos. Cuando esos datos reflejan desigualdades de acceso, registro o calidad de la atención, el modelo tiende a reproducir esas desigualdades en sus recomendaciones. Este problema, ya discutido en el marco teórico en relación con el caso reportado por Obermeyer et al. (2019), exige que la gobernanza institucional trate la composición del conjunto de datos como una cuestión central, y no como un detalle técnico que se delega por completo al equipo de desarrollo.

Una práctica recomendada, presente en varias directrices recientes, es evaluar el desempeño por subgrupo poblacional, como raza, género, edad, región geográfica y nivel socioeconómico, además del desempeño agregado. Un sistema puede mostrar una excelente precisión general y, aun así, tener un desempeño significativamente peor en un subgrupo determinado, algo que pasaría inadvertido en informes que solo presentan promedios. Vayena, Blasimme y Cohen (2018), en un artículo sobre los desafíos éticos del aprendizaje automático en medicina publicado en PLOS Medicine, sostienen que la transparencia sobre la composición del conjunto de datos de entrenamiento, incluidas sus limitaciones de representatividad, debería ser un requisito mínimo para cualquier sistema sometido a aprobación regulatoria o adquisición institucional.

La gobernanza también debe considerar el ciclo de retroalimentación entre el uso de un sistema y la producción de nuevos datos. Si un algoritmo de triaje dirige menos recursos a un grupo determinado, los resultados de ese grupo tienden a empeorar, generando datos que, si se usan para reentrenar el modelo en el futuro, refuerzan el patrón inicial de discriminación. Este ciclo de retroalimentación es particularmente peligroso porque puede ocurrir de manera silenciosa, sin que los gestores adviertan el origen del problema. Documentar el origen de los datos, revisar periódicamente su composición y mantener procesos de auditoría de equidad son medidas concretas para reducir este riesgo.

VALIDACIÓN CLÍNICA CONTEXTUAL

Las métricas de desempeño obtenidas en un entorno de desarrollo, como la sensibilidad, la especificidad y el área bajo la curva ROC calculada sobre un conjunto de prueba retrospectivo, no sustituyen la validación en el contexto real de uso. La diferencia entre el desempeño en condiciones controladas y el desempeño en la práctica clínica cotidiana es un tema recurrente en la literatura, frecuentemente descrito como la brecha entre la validación técnica y la utilidad clínica.

Kelly et al. (2019) revisaron estudios sobre inteligencia artificial en salud y encontraron que la mayoría de ellos no realiza validación externa, es decir, no prueba el modelo en una población o institución distinta de la utilizada en su desarrollo. Sin este paso, no es posible saber si el desempeño reportado se mantiene cuando el sistema se transfiere a un entorno de atención diferente, con un perfil de pacientes distinto, equipos de captura de datos diferentes y rutinas de trabajo diferentes.

La validación contextual también debe evaluar el impacto del sistema sobre las decisiones clínicas reales y los resultados de los pacientes, y no solo la precisión aislada de la predicción. Un modelo puede predecir correctamente el riesgo de una complicación determinada y, aun así, no producir ningún beneficio si la recomendación no cambia la práctica clínica, o si los profesionales no confían lo suficiente en la herramienta como para actuar sobre ella. Los estudios prospectivos que comparan resultados antes y después de la implementación, y las estrategias de implementación gradual, comenzando con unidades piloto antes de expandirse a toda la red, son formas de reducir el riesgo de generalizar prematuramente resultados obtenidos en condiciones artificiales.

TRANSPARENCIA Y EXPLICABILIDAD

Los profesionales de la salud y los pacientes necesitan comprender, hasta cierto punto, el papel que desempeña un sistema de inteligencia artificial en una decisión clínica. Esto no significa que cada usuario deba entender la arquitectura matemática del modelo, sino que la organización debe garantizar explicaciones adecuadas al contexto y a la audiencia, evitando tanto la jerga técnica excesiva como las simplificaciones que inducen confianza excesiva o rechazo infundado.

Amann et al. (2020) proponen distinguir entre diferentes capas de explicabilidad: la explicación técnica dirigida a desarrolladores y auditores, la justificación clínica dirigida a los profesionales de la salud, quienes necesitan entender por qué una recomendación dada tiene sentido dado el

cuadro clínico del paciente, y la comunicación dirigida al paciente, que requiere un lenguaje accesible y atención a las diferencias en alfabetización en salud. La ausencia de cualquiera de estas capas debilita la gobernanza del sistema, incluso si las demás están bien desarrolladas.

La documentación institucional cumple un papel central en esta dimensión. Las fichas técnicas que describen el propósito del sistema, el conjunto de datos utilizado, las limitaciones conocidas, los subgrupos poblacionales para los cuales se evaluó el desempeño y el historial de actualizaciones deben estar disponibles para consulta por parte de los profesionales y, cuando corresponda, de los pacientes y los organismos reguladores. Este tipo de documentación, análoga al prospecto de un medicamento, convierte la transparencia de un principio abstracto en una práctica institucional verificable.

SUPERVISIÓN HUMANA

La presencia de un profesional entre el sistema y la decisión final suele presentarse como una garantía suficiente de seguridad, pero la literatura muestra que este supuesto es frágil. Una supervisión humana efectiva requiere que el profesional cuente con el tiempo, la autonomía y la competencia necesarios para revisar críticamente la recomendación, en lugar de simplemente validarla de manera automática. Cuando la carga de trabajo es excesiva, cuando la organización trata el desacuerdo con el sistema como una desviación que debe sancionarse, o cuando el resultado algorítmico se presenta con una autoridad que desalienta el cuestionamiento, la supervisión se vuelve simbólica.

Cabitza et al. (2017) describen este fenómeno como una forma de complacencia con la automatización, en la cual la presencia humana en el proceso deja de funcionar como una capa de seguridad y pasa a legitimar decisiones que, en la práctica, nunca fueron realmente revisadas. Los protocolos institucionales deben definir explícitamente en qué situaciones una recomendación algorítmica puede aceptarse directamente, en cuáles requiere una revisión obligatoria y en cuáles debe descartarse frente a señales clínicas contradictorias. Estos protocolos también deben exigir que se registre la decisión del profesional, lo que permite evaluar posteriormente si la supervisión humana efectivamente influye en los resultados o si funciona solo como una formalidad.

La capacitación continua es una condición necesaria para que la supervisión humana sea sustantiva. Los profesionales necesitan comprender, en

términos generales, cómo funciona el sistema, cuáles son sus limitaciones conocidas y en qué situaciones su desempeño tiende a ser menos confiable. Sin este nivel mínimo de conocimiento, la autonomía formal para discrepar de una recomendación algorítmica no se traduce en una capacidad real de ejercerla.

IMPUGNACIÓN Y PARTICIPACIÓN DEL PACIENTE

Los pacientes afectados por decisiones respaldadas por sistemas automatizados necesitan acceso a información sobre esta participación y a canales efectivos para plantear preguntas. Este derecho de impugnación cumple una doble función: protege los intereses individuales de los pacientes y funciona como un mecanismo de corrección para la propia organización, ya que las quejas y preguntas a menudo revelan fallas que no aparecen en las métricas de desempeño agregado.

La comunicación sobre el uso de la inteligencia artificial debe tener en cuenta las diferencias en la alfabetización en salud y en la familiaridad con la tecnología entre los distintos grupos de pacientes. Una explicación estandarizada, redactada en lenguaje técnico, puede cumplir los requisitos formales de transparencia sin informar realmente al paciente sobre lo que está en juego. Las iniciativas de participación social, como los consejos de usuarios y las consultas públicas sobre nuevos sistemas antes de su adopción a gran escala, han sido recomendadas por las directrices internacionales como una forma de incorporar esta perspectiva de manera estructurada, y no meramente reactiva.

La Organización Mundial de la Salud (2021), en su documento sobre ética y gobernanza de la inteligencia artificial para la salud, destaca la participación pública como uno de los principios centrales para la adopción responsable de estas tecnologías, argumentando que las decisiones sobre sistemas que afectan directamente la atención de salud no deben tomarse exclusivamente por los desarrolladores de tecnología y los gestores, sin involucrar a las comunidades y a los pacientes afectados.

MONITOREO Y GESTIÓN DE INCIDENTES

El desempeño de un sistema de inteligencia artificial puede deteriorarse después de la implementación, incluso sin ningún cambio en el código del modelo. Los cambios en la población atendida, en los protocolos clínicos, en los equipos utilizados para capturar exámenes o en las prácticas de documentación de las historias clínicas pueden alterar la distribución de los datos de entrada, un fenómeno que la literatura técnica denomina deriva de

datos. Sin un monitoreo continuo, este tipo de deterioro puede pasar inadvertido durante largos períodos.

Sendak et al. (2020) describen, a partir de la experiencia de implementación de un modelo de predicción de sepsis, cómo la etapa de monitoreo posterior a la implementación exigió tanta atención institucional como el desarrollo mismo del algoritmo, incluyendo ajustes en los umbrales de alerta, revisión periódica del desempeño y canales para que los profesionales reportaran los casos en que la recomendación parecía inadecuada. Este tipo de estructura de monitoreo debería ser una parte obligatoria de cualquier proyecto de inteligencia artificial en salud, y no un complemento opcional.

La gestión de incidentes requiere el registro sistemático de los eventos adversos relacionados con el uso del sistema, la investigación de sus causas y la comunicación transparente sobre las medidas correctivas. Cuando un riesgo identificado es grave y no puede mitigarse rápidamente, la suspensión temporal del uso del sistema debe ser una opción institucional disponible y efectivamente ejercida, y no una mera posibilidad teórica escrita en un contrato.

CAPACIDAD INSTITUCIONAL

La gobernanza responsable de la inteligencia artificial en salud depende de una capacidad institucional concreta: equipos multidisciplinarios capaces de evaluar propuestas técnicas, infraestructura de datos adecuada, contratos bien redactados y un liderazgo comprometido con la supervisión continua de los sistemas adoptados. Las organizaciones necesitan desarrollar la competencia para evaluar críticamente a los proveedores, lo que incluye la capacidad de solicitar e interpretar evidencia de validación, negociar cláusulas contractuales sobre auditoría y desempeño, y capacitar a los profesionales para el uso adecuado de la tecnología.

Esta capacidad no se distribuye de manera equitativa entre las organizaciones de salud. Las redes con pocos recursos, especialmente en regiones con menor desarrollo económico, dependen casi por completo de proveedores externos para evaluar y ajustar los sistemas de inteligencia artificial, lo que reduce su autonomía de decisión y su poder de negociación. Sun y Medaglia (2019), en su estudio sobre la adopción de la inteligencia artificial en hospitales chinos, encontraron que las barreras organizacionales, y no solo las limitaciones técnicas, fueron decisivas para explicar el éxito o el fracaso de las diferentes iniciativas, reforzando la idea

de que la inversión en capacidad institucional es tan importante como la inversión en la propia tecnología.

Las políticas públicas de apoyo, incluidos los estándares técnicos compartidos, la infraestructura pública para la evaluación independiente y los programas de capacitación, pueden reducir esta asimetría y evitar que la inteligencia artificial en salud se convierta en un factor más que concentre ventajas entre las instituciones que ya cuentan con más recursos.

MARCO PROPUESTO

El marco desarrollado en este artículo reúne las seis dimensiones discutidas: propósito clínico legítimo, calidad y representatividad de los datos, validación contextual, transparencia y explicabilidad, supervisión humana efectiva y mecanismos de impugnación, y monitoreo continuo, que atraviesa todas las demás. Estas dimensiones no deben entenderse como pasos secuenciales y aislados, sino como aspectos que se retroalimentan entre sí a lo largo de todo el ciclo de vida del sistema, desde la adquisición hasta su eventual retiro.

La decisión de adoptar un sistema de inteligencia artificial en salud debe tratarse como revisable desde el inicio, y no como un compromiso definitivo asumido en el momento de la adquisición. Esto implica mantener registros que permitan comparar el desempeño previsto con el desempeño observado, revisar periódicamente la vigencia de la tecnología a la luz de las alternativas disponibles y, cuando sea necesario, discontinuar su uso. Esta orientación hacia el valor público, en lugar de hacia la eficiencia técnica aislada, es el hilo que conecta las seis dimensiones del modelo.

ADQUISICIÓN, REGULACIÓN Y RENDICIÓN DE CUENTAS

La adquisición de sistemas de inteligencia artificial en salud requiere criterios que combinen la evaluación clínica, técnica y legal. Las licitaciones y los contratos deben especificar requisitos mínimos de desempeño, documentación técnica exigida, estándares de seguridad de la información, requisitos de interoperabilidad con los sistemas existentes y condiciones de soporte técnico continuo. Sin estos elementos, la organización compradora queda en desventaja informativa frente al proveedor, sin poder evaluar adecuadamente si el producto entregado corresponde a lo prometido.

La asignación de responsabilidades entre los desarrolladores de tecnología, las instituciones de salud y los profesionales que utilizan el sistema debe ser explícita. La multiplicación de actores involucrados en el desarrollo y uso de

un sistema de inteligencia artificial no debe generar vacíos de rendición de cuentas, es decir, situaciones en las que, al ocurrir un daño, cada parte alega que la falla es de otra. Los protocolos de investigación de incidentes, definidos de antemano en lugar de improvisados después de que ocurre un problema, ayudan a esclarecer las causas y a asignar la responsabilidad de manera justa.

La regulación externa y la gobernanza institucional interna son complementarias, no sustitutas entre sí. El cumplimiento formal de los requisitos regulatorios, como el registro ante una autoridad sanitaria o la certificación de calidad, no elimina la necesidad de una evaluación local sobre la idoneidad del sistema para el contexto específico de uso. Las organizaciones necesitan dar seguimiento a las actualizaciones regulatorias y a los cambios en la forma en que el sistema se utiliza realmente en la práctica, la cual puede alejarse de su uso original previsto y aprobado.

Los contratos también deben prever mecanismos de auditoría independiente, portabilidad de los datos en caso de cambio de proveedor y condiciones claras para la terminación de la relación contractual. La dependencia excesiva de un único proveedor, sin cláusulas que garanticen la continuidad y la transparencia, debilita la capacidad de la institución para ejercer un control efectivo sobre la tecnología que ha adoptado.

CAPACITACIÓN PROFESIONAL Y CULTURA DE SEGURIDAD

Los profesionales de la salud necesitan comprender, incluso a un nivel no técnico, las capacidades y los límites de los sistemas de inteligencia artificial con los que trabajan. Los programas de capacitación deben incluir nociones básicas sobre cómo interpretar los resultados algorítmicos, qué tipos de sesgo pueden afectar las recomendaciones, principios de privacidad de los datos clínicos y estrategias de comunicación con los pacientes sobre el papel de la tecnología en la atención que reciben. El objetivo no es convertir a cada profesional en un especialista en ciencia de datos, sino desarrollar un juicio crítico suficiente para usar la herramienta con seguridad.

Una cultura organizacional que permita reportar errores y desacuerdos sin temor a sanciones es una condición necesaria para que el monitoreo continuo funcione en la práctica. Los entornos punitivos tienden a reducir el reporte espontáneo de problemas, ocultando las fallas hasta que se vuelven graves. Los canales de denuncia confidenciales y el análisis sistémico de los incidentes reportados, en lugar de la culpa individual inmediata, fortalecen la capacidad de aprendizaje de la organización.

Los equipos multidisciplinarios, que reúnen conocimientos clínicos, técnicos y éticos, ayudan a identificar problemas que pasarían inadvertidos en análisis puramente técnicos. Las reuniones periódicas de revisión del desempeño, con la participación de diferentes categorías profesionales, permiten dar seguimiento tanto a indicadores cuantitativos como a informes cualitativos sobre la experiencia de uso del sistema.

Los pacientes y las comunidades también necesitan información accesible sobre cómo se utilizan los sistemas de inteligencia artificial en su atención. Los materiales de comunicación redactados en lenguaje sencillo y probados con usuarios reales antes de su publicación ayudan a construir una comprensión de los usos y derechos, contribuyendo a la confianza institucional de manera más consistente que las declaraciones genéricas sobre innovación tecnológica.

INTEROPERABILIDAD Y CONTINUIDAD DE LA ATENCIÓN

Los sistemas de inteligencia artificial dependen de datos que fluyen con frecuencia entre distintos servicios y niveles de atención. La interoperabilidad técnica, que garantiza que diferentes sistemas puedan intercambiar datos, y la interoperabilidad semántica, que garantiza que esos datos conserven el mismo significado clínico al transferirse entre sistemas, son condiciones esenciales para evitar la fragmentación de la información y, con ella, una pérdida de calidad en las recomendaciones generadas.

La continuidad de la atención puede verse comprometida cuando las recomendaciones generadas por un sistema en un servicio no acompañan al paciente a lo largo de su trayectoria asistencial. Una alerta de riesgo generada durante una internación hospitalaria, por ejemplo, pierde su utilidad si no se comunica al equipo de atención primaria responsable del seguimiento tras el alta. La integración entre los sistemas de inteligencia artificial y las historias clínicas electrónicas debe planificarse desde el inicio del proyecto, y no tratarse como un asunto secundario.

Los cambios de proveedor o de versión del sistema requieren procesos de migración de datos que preserven el historial clínico relevante. Los contratos de adquisición deben prever explícitamente condiciones de portabilidad y preservación de registros, evitando que un cambio de tecnología resulte en la pérdida de información asistencial importante. Una interoperabilidad bien estructurada también facilita los procesos de auditoría externa y la investigación científica sobre el desempeño real de los

sistemas en uso, siempre que se respeten los requisitos de privacidad y de finalidad legítima en el uso secundario de los datos.

EVALUACIÓN ECONÓMICA Y VALOR PÚBLICO

La decisión de implementar un sistema de inteligencia artificial en salud debe considerar su costo total, que incluye no solo la licencia o la adquisición inicial, sino también la infraestructura de datos, la capacitación profesional, el mantenimiento técnico continuo y el monitoreo del desempeño a lo largo del tiempo. Las proyecciones de ahorro presentadas por los proveedores a menudo no se materializan cuando se mantienen procesos manuales paralelos como medida de precaución, o cuando la implementación requiere inversiones complementarias no previstas inicialmente.

La evaluación económica de los sistemas de inteligencia artificial en salud debe incorporar dimensiones de seguridad, equidad y calidad de los resultados, y no solo indicadores de eficiencia operativa. Una reducción en el tiempo de atención, por ejemplo, no representa un beneficio real cuando se logra a costa de una mayor tasa de error diagnóstico o de barreras adicionales de acceso para determinados grupos de pacientes.

Los proyectos piloto, realizados a escala reducida antes de la expansión a toda la red de atención, ayudan a producir estimaciones más confiables de los costos reales y de los efectos clínicos y organizacionales que las proyecciones basadas exclusivamente en datos proporcionados por el desarrollador de la tecnología. Los criterios objetivos para decidir la ampliación de escala, definidos antes de iniciar el piloto, evitan que las decisiones de expansión estén impulsadas por presión comercial o por entusiasmo tecnológico sin fundamento en evidencia local.

La transparencia sobre los costos y resultados de los proyectos de inteligencia artificial en salud permite el control social y la priorización adecuada de los recursos públicos. Dado que los recursos destinados a la salud implican decisiones con un alto costo de oportunidad, las inversiones en tecnología deben justificarse públicamente con el mismo rigor aplicado a otras decisiones sobre la asignación de recursos asistenciales.

AGENDA DE INVESTIGACIÓN Y POLÍTICAS

Las investigaciones futuras sobre la gobernanza de la inteligencia artificial en salud deberían priorizar estudios realizados en entornos reales de práctica clínica, con poblaciones diversas, en lugar de basarse

exclusivamente en evaluaciones retrospectivas con bases de datos controladas. Los estudios prospectivos, que dan seguimiento a los resultados clínicos antes y después de la implementación de un sistema, ofrecen evidencia más robusta de un beneficio real que las comparaciones puramente técnicas entre modelos.

Es necesario profundizar la comprensión sobre cómo interactúan las recomendaciones algorítmicas con el juicio clínico profesional en condiciones reales de trabajo, incluyendo factores organizacionales como la carga de trabajo, la cultura institucional y la relación de confianza entre el equipo y la tecnología. Estos factores a menudo explican variaciones en el desempeño y en la adherencia que no pueden predecirse únicamente a partir de las características técnicas del modelo.

Las políticas públicas tienen un papel importante que desempeñar en el apoyo a estándares técnicos comunes, infraestructura compartida para la evaluación independiente y mecanismos de certificación que reduzcan la asimetría de información entre los proveedores y las organizaciones de salud, especialmente las más pequeñas. Las redes más pequeñas, con recursos limitados, necesitan apoyo estructurado para participar en la adopción responsable de estas tecnologías sin depender por completo de la buena voluntad de los proveedores comerciales.

Finalmente, la gobernanza de la inteligencia artificial en salud debe entenderse como un proceso en constante evolución, que se ajusta tanto al avance tecnológico como a la acumulación de evidencia sobre sus efectos reales en la práctica clínica. Las revisiones periódicas de las directrices institucionales y regulatorias son necesarias para mantener actualizado el marco de gobernanza frente a cambios que ocurren a un ritmo más rápido que los ciclos habituales de revisión regulatoria.

DISCUSIÓN AMPLIADA E IMPLICACIONES INSTITUCIONALES

El análisis desarrollado a lo largo de este artículo indica que los cambios tecnológicos producen resultados consistentes solo cuando se acompañan de un diseño institucional adecuado. La disponibilidad de herramientas, datos y habilidades individuales no sustituye a las reglas de coordinación, la definición de la autoridad para tomar decisiones y los mecanismos de rendición de cuentas. Las organizaciones de salud necesitan hacer explícitos sus propósitos, asignar responsabilidades y definir criterios de evaluación verificables. Esta estructura no elimina la incertidumbre, pero crea las condiciones para que las decisiones sean justificadas, revisadas y mejoradas

a lo largo del tiempo, en lugar de permanecer implícitas en las elecciones informales de gestores o proveedores.

Un segundo aspecto relevante se refiere a la relación entre la capacidad técnica y la legitimidad institucional. Los proyectos de inteligencia artificial en salud pueden mostrar un buen desempeño operativo, medido por métricas técnicas, y aun así enfrentar resistencia por parte de los profesionales y desconfianza por parte de los pacientes cuando los afectados no comprenden los objetivos o carecen de canales reales para participar en las decisiones sobre la adopción de la tecnología. La legitimidad se construye mediante la combinación de evidencia técnica, transparencia procedimental y trato justo hacia los distintos grupos afectados. Los procesos deliberativos no eliminan los conflictos de interés ni de valores, pero ayudan a poner de manifiesto prioridades y consecuencias que los análisis puramente técnicos tienden a pasar por alto.

La distribución desigual de recursos entre las organizaciones de salud condiciona directamente la gobernanza de estas tecnologías. Las instituciones más grandes cuentan con equipos especializados, infraestructura robusta y mayor poder de negociación frente a los proveedores, mientras que las organizaciones más pequeñas dependen casi por completo del apoyo externo. Las políticas de cooperación entre instituciones y de intercambio de capacidades de evaluación pueden reducir esta brecha. Toda evaluación de impacto de un sistema de inteligencia artificial en salud debería examinar no solo los resultados agregados, sino también si la innovación amplía o reduce las asimetrías entre territorios, grupos poblacionales y distintas unidades dentro de una misma red de atención.

La proporcionalidad entre riesgo y control, discutida en la sección sobre el propósito clínico, tiene implicaciones prácticas importantes para el diseño de los procesos institucionales. Las aplicaciones de bajo impacto, como las herramientas de agenda administrativa, pueden gestionarse con procedimientos relativamente simples. Las decisiones que afectan directamente el diagnóstico, el tratamiento o la priorización del acceso a recursos escasos requieren documentación detallada, revisión por parte de organismos independientes y mecanismos de apelación claros para los pacientes que se sientan perjudicados. Un sistema de gobernanza que trate toda aplicación con el mismo grado de rigor tiende a producir dos problemas opuestos: burocracia excesiva para los casos de bajo riesgo y control insuficiente para los casos de alto impacto.

La capacidad de aprendizaje institucional es otro componente decisivo para la sostenibilidad de la gobernanza propuesta. Los proyectos de inteligencia artificial en salud deben tratarse como procesos adaptativos, formulados con hipótesis explícitas, indicadores de monitoreo y momentos formales de revisión, en lugar de implementaciones definitivas evaluadas únicamente en su lanzamiento. Los resultados inesperados, positivos o negativos, deben generar ajustes concretos en los modelos, los contratos y las rutinas de trabajo. Las organizaciones que ocultan fallas, por temor a rendir cuentas o a sufrir daño reputacional, pierden oportunidades valiosas de corrección y tienden a repetir los mismos problemas en proyectos futuros. Los entornos que combinan la rendición de cuentas con un margen legítimo para corregir el rumbo favorecen este tipo de aprendizaje.

Desde el punto de vista gerencial, la gobernanza de la inteligencia artificial en salud requiere equipos genuinamente multidisciplinarios, en los que el conocimiento técnico dialogue con la experiencia clínica operativa, la formación jurídica, la reflexión ética y la capacidad de comunicarse con pacientes y comunidades. Esta diversidad de conocimientos reduce los puntos ciegos que tienden a aparecer cuando las decisiones se concentran en un único grupo profesional, ya sea técnico, clínico o administrativo. Al mismo tiempo, los equipos multidisciplinarios necesitan un lenguaje común y procesos de decisión bien definidos, o corren el riesgo de fragmentarse y producir decisiones contradictorias entre distintas áreas de la misma organización.

La sostenibilidad financiera y organizacional de los proyectos de inteligencia artificial en salud depende de recursos recurrentes, y no solo de la inversión inicial de implementación. Las iniciativas financiadas únicamente para la fase de lanzamiento tienden a deteriorarse una vez que finaliza el período del proyecto o el financiamiento externo. Los costos de mantenimiento, las actualizaciones del modelo, la capacitación profesional continua y el monitoreo del desempeño deben incluirse desde la etapa inicial de planificación presupuestaria. La decisión de continuar, ajustar o discontinuar una iniciativa de inteligencia artificial en salud debe considerar explícitamente el valor efectivamente producido, los riesgos identificados durante el uso y las alternativas disponibles, incluida la posibilidad de retornar a procesos no automatizados cuando estos resulten más adecuados al contexto.

Finalmente, el desarrollo de los sistemas de inteligencia artificial en salud debe mantenerse orientado hacia el propósito público y organizacional, y no hacia la sofisticación tecnológica como fin en sí mismo. El valor de cualquier

herramienta se deriva de su capacidad para resolver problemas concretos, fortalecer la capacidad de las instituciones de salud y producir beneficios distribuidos de manera justa entre la población atendida. La evaluación de cualquier proyecto debería regresar periódicamente a la pregunta que guio su creación: qué necesidades se atendieron realmente, para quién, y con qué consecuencias para los distintos grupos afectados.

LIMITACIONES DEL MODELO Y CRITERIOS DE APLICACIÓN

El modelo presentado en este artículo tiene un carácter teórico y debe adaptarse al contexto específico de cada organización. Los servicios de salud varían sustancialmente en tamaño, misión institucional, disponibilidad de recursos, entorno regulatorio y madurez tecnológica. Aplicar de manera mecánica las recomendaciones aquí discutidas, sin esta adaptación, puede crear nuevos problemas en lugar de resolverlos. Antes de implementar cualquiera de las dimensiones propuestas, es necesario mapear los procesos existentes, identificar a las partes interesadas relevantes, evaluar los riesgos específicos del contexto local y reconocer honestamente las capacidades disponibles.

Otra limitación se deriva de la velocidad del cambio tecnológico en el campo de la inteligencia artificial. Las herramientas, los estándares técnicos y las configuraciones del mercado evolucionan rápidamente, mientras que las reglas institucionales y los procesos regulatorios suelen tardar más en consolidarse. La gobernanza necesita ser lo suficientemente estable como para brindar certeza jurídica y operativa, pero lo suficientemente flexible como para incorporar nueva evidencia sin requerir una revisión completa con cada avance tecnológico. Las revisiones programadas a intervalos regulares, en lugar de solo en respuesta a crisis, ayudan a equilibrar estas dos necesidades aparentemente contradictorias.

La disponibilidad y calidad de los datos utilizados para la evaluación y el monitoreo también limitan, en la práctica, la forma en que puede aplicarse el modelo. Los indicadores de desempeño pueden estar incompletos, mostrar inconsistencias entre distintas fuentes o ser producidos directamente por los propios proveedores de tecnología, lo que compromete su independencia. Triangular evidencia de diferentes fuentes, incluidas auditorías externas y observaciones cualitativas del uso cotidiano del sistema, reduce la dependencia de una única medida de desempeño. Cuando el grado de incertidumbre sobre el funcionamiento real de un sistema es alto, las decisiones de ampliar su uso deben ser graduales y reversibles, evitando compromisos institucionales difíciles de deshacer.

Los conflictos de interés merecen atención específica en cualquier aplicación práctica del modelo. Los proveedores de tecnología pueden influir, directa o indirectamente, en la definición de los criterios de éxito y presentar los resultados de manera selectiva, destacando escenarios favorables y omitiendo limitaciones relevantes. Los consultores contratados, los investigadores vinculados a los proyectos y los gestores que buscan resultados rápidos también pueden tener incentivos que no se alinean plenamente con los intereses de los pacientes. Las declaraciones formales de conflicto de interés, la transparencia sobre los términos contractuales y la revisión por partes independientes ayudan a mitigar, aunque no a eliminar por completo, este tipo de sesgo institucional.

La aplicación efectiva del modelo requiere un liderazgo institucional comprometido. Sin el apoyo explícito de la alta dirección, los comités de evaluación y los protocolos de supervisión tienden a convertirse en formalidades burocráticas con poco efecto práctico sobre las decisiones reales de la organización. Los líderes deben poner a disposición recursos suficientes, responder de manera sustantiva a las recomendaciones producidas por estos procesos y aceptar, cuando sea necesario, la posibilidad de detener iniciativas en curso, incluso frente a inversiones ya realizadas.

Los criterios prácticos para aplicar el modelo deberían incluir la relevancia clínica u organizacional de la tecnología propuesta, la proporcionalidad entre el riesgo y el nivel de control requerido, la capacidad institucional disponible, la participación adecuada de los profesionales y pacientes afectados, y una estructura de monitoreo continuo. Cada uno de estos criterios puede registrarse en una matriz de decisión simple, lo que facilita la comparación entre alternativas tecnológicas competidoras y mantiene un registro de las justificaciones utilizadas para cada elección, creando un historial útil para evaluaciones futuras.

Finalmente, cabe señalar que la decisión de no adoptar una tecnología determinada, o de discontinuar su uso tras un período de evaluación, es un resultado legítimo y, en ocasiones, deseable del proceso de gobernanza. La innovación responsable en salud incluye la capacidad institucional de reconocer cuándo los beneficios esperados de un sistema no compensan sus costos y riesgos reales. Los procesos de discontinuación bien planificados deben preservar los datos clínicos relevantes, garantizar la continuidad de los servicios prestados a los pacientes y mantener un registro del conocimiento acumulado durante el período de uso, de modo que esa experiencia pueda informar decisiones futuras. Estas limitaciones no

disminuyen la utilidad del marco propuesto en este artículo; por el contrario, indican las condiciones necesarias para su uso prudente. La principal contribución del modelo consiste en organizar, de manera integrada, cuestiones que a menudo son abordadas de forma fragmentada por diferentes áreas dentro de una misma organización de salud.

CONSIDERACIONES FINALES

La inteligencia artificial puede aportar una contribución significativa a los sistemas de salud, apoyando el diagnóstico, el triaje, la gestión de recursos y la planificación de la atención. A lo largo de este artículo se ha argumentado que estos beneficios dependen de estructuras de gobernanza capaces de traducir principios éticos amplios en responsabilidades concretas, procedimientos verificables y mecanismos de corrección efectivos. El desempeño técnico medido de forma aislada, mediante métricas estadísticas de precisión o discriminación, no garantiza la equidad en la atención, la seguridad del paciente ni el fortalecimiento de la capacidad institucional de las organizaciones de salud involucradas.

El modelo propuesto, organizado en torno a seis dimensiones interrelacionadas —propósito clínico legítimo, calidad y representatividad de los datos, validación contextual, transparencia y explicabilidad, supervisión humana efectiva, mecanismos de impugnación y monitoreo continuo—, busca ofrecer directrices aplicables tanto a los gestores de servicios de salud como a los formuladores de políticas responsables de regular estas tecnologías. La contribución central del artículo consiste en tratar estos requisitos como componentes de un único sistema, y no como listas de verificación independientes que deben cumplirse de forma aislada.

Se reconoce, no obstante, que el modelo tiene un carácter teórico y que su aplicación práctica requiere una adaptación cuidadosa a cada contexto institucional específico. Investigaciones futuras podrían poner a prueba la aplicabilidad del marco en diferentes redes de atención, evaluando empíricamente si las organizaciones que adoptan estas seis dimensiones logran efectivamente mejores resultados en términos de seguridad, equidad y legitimidad institucional que las organizaciones que tratan a la inteligencia artificial predominantemente como una cuestión técnica. Los estudios comparativos entre sistemas de salud con diferentes niveles de capacidad institucional y diferentes marcos regulatorios también podrían aclarar en qué medida los principios discutidos en este artículo se sostienen en contextos distintos de aquel que originalmente motivó su formulación.

Finalmente, vale la pena reafirmar que la incorporación responsable de la inteligencia artificial a los sistemas de salud no es meramente un desafío técnico que deban resolver ingenieros y científicos de datos, ni tampoco una cuestión regulatoria que deban resolver los legisladores. Es un proceso continuo de construcción institucional, que requiere coordinación entre distintos actores, deliberación explícita sobre valores en competencia y disposición a revisar las decisiones a la luz de nueva evidencia. La gobernanza discutida a lo largo de este artículo representa un intento de organizar esta complejidad en un marco coherente, sin pretender agotar un debate que, dada la velocidad de las transformaciones tecnológicas en curso, seguirá desarrollándose durante muchos años más.

Una cuestión adicional, poco explorada en la literatura brasileña sobre el tema, se refiere a la relación entre la gobernanza de la inteligencia artificial y la reforma más amplia de los sistemas de salud. En el contexto de los sistemas públicos universales, como el Sistema Único de Salud de Brasil (Sistema Único de Saúde), las decisiones sobre la adopción de tecnología no pueden tomarse de manera aislada por cada unidad de atención, so pena de producir un mosaico de prácticas mutuamente incompatibles, con estándares de calidad y seguridad muy variables de una región a otra. La coordinación entre los niveles de gestión —municipal, estatal y federal— se vuelve necesaria para garantizar que se observen principios mínimos de gobernanza en toda la red, sin que ello impida, sin embargo, adaptaciones locales justificadas por diferencias genuinas en el contexto asistencial.

Esta tensión entre la estandarización nacional y la adaptación local aparece de manera recurrente en otras áreas de la política de salud y no es exclusiva de la inteligencia artificial. La experiencia acumulada con los protocolos clínicos, los sistemas de información en salud y las políticas de regulación del acceso sugiere que las soluciones exitosas tienden a combinar directrices generales definidas a nivel central con margen para el ajuste operativo a nivel local. Aplicada a la inteligencia artificial, esta lógica sugiere que las agencias reguladoras y los órganos de gestión del sistema de salud deberían definir requisitos mínimos de seguridad, transparencia y equidad, dejando a cada organización la tarea de traducir estos requisitos en procesos compatibles con su realidad operativa.

También vale la pena considerar el papel de la investigación académica en el sostenimiento de este proceso de gobernanza. Las universidades y los centros de investigación pueden contribuir de manera significativa, no solo desarrollando nuevos algoritmos, sino también produciendo evaluaciones independientes del desempeño real de los sistemas ya en uso, algo que los

proveedores comerciales rara vez tienen incentivos para hacer de manera imparcial. Las alianzas entre instituciones de investigación y servicios de salud, con salvaguardas adecuadas para la protección de datos y los conflictos de interés, pueden funcionar como un mecanismo complementario de control de calidad, particularmente en sistemas de salud con capacidad regulatoria limitada.

Finalmente, vale la pena señalar la dimensión temporal de la gobernanza discutida en este artículo. Muchas de las fallas reportadas en la literatura sobre inteligencia artificial en salud no ocurrieron en el momento de la implementación inicial, sino que surgieron meses o años después, a medida que el contexto de uso se fue alejando gradualmente de las condiciones originales de validación. Esto refuerza la idea de que la gobernanza no es un evento único ubicado al inicio de un proyecto, sino un proceso continuo que acompaña al sistema durante toda su vida útil. Las organizaciones que tratan la aprobación inicial como el punto final del proceso de evaluación tienden a verse sorprendidas por problemas que podrían haberse identificado y corregido a tiempo, de haber existido una estructura de monitoreo permanente.

REFERENCIAS

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20(1), 310.

Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. JAMA, 318(6), 517-518.

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care: addressing ethical challenges. New England Journal of Medicine, 378(11), 981-983.

Coiera, E. (2007). Putting the technical back into socio-technical systems research. International Journal of Medical Informatics, 76, S98-S103.

Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., & Socher, R. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24-29.

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17(1), 195.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

Panch, T., Mattie, H., & Atun, R. (2019). Artificial intelligence and algorithmic bias: implications for health systems. Journal of Global Health, 9(2), 010318.

Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. Nature Medicine, 25(1), 37-43.

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine, 169(12), 866-872.

Sendak, M. P., Ratliff, W., Sarro, D., Alderton, E., Futoma, J., Gao, M., Nichols, M., Revoir, M., Yashar, F., Miller, C., Kester, K., Sandhu, S., Corey, K., Brajer, N., Tan, C., Lin, A., Brown, T., Engelbosch, S., Anstrom, K., ... Balu, S. (2020). Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. NPJ Digital Medicine, 3(1), 84.

Sun, T. Q., & Medaglia, R. (2019). Mapping the challenges of artificial intelligence in the public sector: evidence from public healthcare. Government Information Quarterly, 36(2), 368-383.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56.

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: addressing ethical challenges. PLOS Medicine, 15(11), e1002689.

Verghese, A., Shah, N. H., & Harrington, R. A. (2018). What this computer needs is a physician: humanism and artificial intelligence. JAMA, 319(1), 19-20.

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2019). Do no harm: a roadmap for responsible machine learning for health care. Nature Medicine, 25(9), 1337-1340.

World Health Organization. (2021). Ethics and governance of artificial intelligence for health. Geneva: World Health Organization.

CONTRIBUCIÓN DEL AUTOR (TAXONOMÍA CREDIT)

Antonio Pires Barbosa: Conceptualización; Metodología; Investigación; Análisis formal; Curación de datos – no aplica, ya que no se utilizaron datos empíricos; Redacción – borrador original; Redacción – revisión y edición; Supervisión; Obtención de financiamiento – no aplica, sin financiamiento externo. Entrada simulada para pruebas técnicas de la plataforma OJS.

DECLARACIONES EDITORIALES

Financiamiento: no se recibió financiamiento externo.

Conflicto de intereses: no aplica.

Uso de inteligencia artificial: no aplica.

DISPONIBILIDAD DE LOS DATOS: NO APLICA; NO SE UTILIZARON DATOS EMPÍRICOS.