INTELIGENCIA ARTIFICIAL Y MODELOS FUNDACIONALES, LA NUEVA ERA EN PRODUCTOS SANITARIOS: REGULACIÓN, VALIDACIÓN Y MONITORIZACIÓN

La Inteligencia Artificial (IA) ya forma parte integral de nuestras vidas y no es ciencia ficción. Desde la utilización de asistentes virtuales personales para organizar nuestra jornada laboral, viajar en vehículos autónomos, hasta en el ámbito sanitario, incluyendo la detección y diagnóstico precoz de enfermedades; la identificación de nuevas observaciones o patrones en la fisiología humana; el desarrollo de diagnósticos y tratamientos personalizados, entre otros, con el objetivo de mejorar la experiencia tanto del usuario como del paciente, la IA es una realidad tangible.

ILHAM BEN YAHYA,
Farmacéutica, Posgrado en Investigación en desarrollo, innovación de medicamentos (Universidad de Navarra). Becario de QA/RA en A3Z advanced S.L.

MIGUEL A. CAMPANERO MARTINEZ,
Doctor en Farmacia, Especialista en Farmacia Industrial y Galénica, Director Técnico A3Z advanced S.L.

En los últimos años, los modelos fundacionales de lenguaje (*Large Language Models, LLM*), como ChatGPT, han irrumpido con fuerza en el ámbito de la salud. Su capacidad para comprender lenguaje natural, generar respuestas contextualizadas y analizar grandes volúmenes de información los posiciona como herramientas potenciales de apoyo en diagnóstico clínico, generación de informes médicos, priorización de casos o asistencia en decisiones clínicas. Sin embargo, su integración en aplicaciones (APPs) de soporte clínico plantea enormes desafíos regulatorios. La naturaleza de estos modelos, entrenados sobre datos masivos y basados en el aprendizaje continuo, la diferencia de otras tecnologías médicas y exige marcos robustos para garantizar su seguridad, eficacia y calidad [1].

Este artículo analiza los requisitos regulatorios fundamentales para la validación y monitorización del correcto funcionamiento de APPs que incorporan LLM en contextos clínicos, con referencia a marcos legales europeos (Reglamentos (UE) 2017/745-MDR [2], 2017/746-IVDR [3], 2024/1689-IA Act [4], estadounidense (FDA) [5-7] y otras guías internacionales relevantes [8,9].

Característica	Modelos tradicionales	Modelos fundacionales
Propósito inicial	Tarea clínica específica	Uso general adaptable
Datos de entrenamiento	Datos clínicos etiquetados	Datos masivos, variados y no estructurados
Adaptabilidad funcional	Limitada: un modelo por tarea	Alta: un modelo puede adaptarse a múltiples tareas
Explicabilidad	Alta: modelos interpretables y transparentes	Limitada: modelo complejo tipo "caja negra"
Evolución tras despliegue	Modelo estático	Prevista: requiere planificación y control continuado (PCCP)
Riesgo regulatorio	Bajo a moderado por su previsibilidad	Alto por sesgos, dificultad de validación)
Reentrenamiento	Necesario para cada nueva tarea	No siempre necesario (finetuning)
Ejemplo de uso	Reconocimiento facial	Asistente clínico digital

Tabla 1. Diferencias clave entre modelo tradicional y modelo fundacional.

¿Qué se entiende por modelo fundacional?

Un modelo fundacional es un tipo de modelo de inteligencia artificial de propósito general entrenado a gran escala con datos masivos y diversos, tanto mediante aprendizaje no supervisado, lo que le permite adquirir representaciones profundas del lenguaje, imágenes, código u otros dominios, lo que da como resultado que el modelo pueda ser posteriormente adaptado a una amplia gama de diferentes usos previstos. Por tanto, y desde un punto de vista regulatorio, un modelo fundacional no puede ser considerado como un producto sanitario ya que tiene un propósito general. Mas bien, debe ser considerado como un elemento que va a utilizarse en el diseño y desarrollo de un producto sanitario con un uso previsto determinado, y que debe ser previamente homologado para asegurar que cumple los requisitos generales de seguridad y funcionamiento relativos al diseño requeridos por el MDR y el IVDR. A continuación, se presenta una tabla comparativa (Tabla 1) que resume las principales diferencias entre los modelos tradicionales de inteligencia artificial y los modelos de IA de nueva generación.

Marco regulatorio aplicable

En Europa, los productos sanitarios que incorporan algoritmos basados en inteligencia artificial se regulan bajo el MDR y el IVDR, reglamentos que incluyen dentro de su alcance el producto software con propósito médico específico (SaMD, Software as a Medical Device). El AI Act añade una capa específica de regulación para sistemas de IA de alto riesgo, entre los que se incluyen los productos que deban ser certificados previamente a su puesta en servicio. Este Reglamento aplica también a otras aplicaciones que incorporan inteligencia artificial que podemos encontrar en el entorno sanitario pero que no cumplen la definición de producto sanitario. En la figura 1, se muestra un árbol de decisión útil para decidir el marco regulatorio que aplica en cada caso.

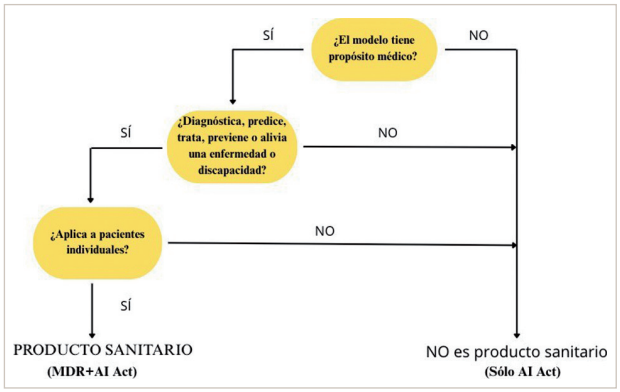

Figura 1. Árbol de decisión regulatoria de PS.

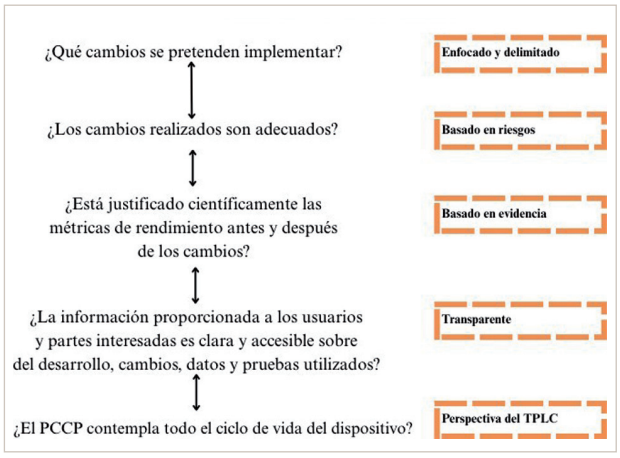

Figura 2. Esquema del proceso para la implementación de cambios predeterminados.

Por su parte, en Estados Unidos, la FDA ha desarrollado un marco específico para SaMD basados en aprendizaje automático, que introduce el concepto de Planes

Fase 1 Validación del diseño		
Validación de datos		
Garantizar la calidad	**Garantizar trazabilidad**	**Garantizar confidencialidad**
Para garantizar la consistencia e integridad de los datos según los principios ALCOA (Atribuible, Legible, Contemporáneo, Original, Accurate-exacto-)	Demostrar en cualquier momento y por escrito todo lo que se ha hecho con los datos, desde que se recogen hasta que se usan para entrenar o validar un sistema de IA.	Para demostrar la confidencialidad de los datos en un sistema de IA, se tiene que dejar constancia documental del cumplimiento con el Reglamento General de Protección de Datos (RGPD), y que los datos se han tratado de forma segura y con base legal clara.
Validación de algoritmos		
Para garantizar su robustez técnica y asegurar la fiabilidad, seguridad del sistema.		
Validación de outputs		
Para garantizar su utilidad clínica y aceptación por parte de los profesionales. Los outputs pueden incluir recomendaciones diagnósticas, predicciones de riesgos, alertas o decisiones automatizadas.		

Figura 3. Validación del diseño de PS con IA.

de Control de Cambios Predeterminado (PCCP) [6], que permiten planificar y gestionar cambios adaptativos sin tener que reiniciar todo el proceso de aprobación tras cada actualización. Europa ha adoptado recientemente este enfoque [10], permitiendo que los cambios predeterminados sean evaluados previamente a evaluación de la conformidad. La figura 2, muestra el proceso que se debe seguir para implementar estos cambios:

Los tres marcos normativos MDR, IVDR y el IA Act destacan la importancia de realizar pruebas en diversas condiciones, documentar los procesos de evaluación y establecer una monitorización continua para asegurar el cumplimiento. Estas regulaciones obligan a los fabricantes a aportar evidencias que respalden sus afirmaciones, garantizando que los productos sanitarios que incorporan inteligencia artificial (MDAI) se adhieran a estándares estrictos de seguridad y rendimiento. Esto incluye la documentación de las pruebas, de los procesos de validación y la monitorización continua para abordar cualquier problema que pueda surgir. Las regulaciones ofrecen flexibilidad en la elección de los métodos de validación, pero subrayan la importancia de demostrar que el MDAI funciona correctamente y cumple los requisitos especificados.

Sistemática de validación

La validación de un producto sanitario que incorpora IA debe abordarse de manera estructurada y por fases. En este contexto, la validación constituye un eje fundamental, y se divide en dos etapas principales: validación analítica o técnica, que se realiza antes del despligue, durante las fases de diseño y desarrollo; y la validación clínica, que tiene lugar después del despliegue, cuando el producto ya está instalado y se encuentra en entorno de uso real o simulado.

La validación técnica implica la aplicación sistemática de métodos de evaluación de la calidad de los datos y del modelo para detectar posibles modos de fallo en su comportamiento. Esto incluye métricas orientadas al modelo, como el rendimiento predictivo, la robustez o la interpretabilidad. Asimismo, se integran métricas orientadas a los datos relacionadas con la determinación del tamaño muestral, la dispersión, o el sesgo. Un análisis estadístico riguroso de estas métricas constituye un punto crítico en investigación y desarrollo industrial, y es por ello una pieza clave dentro del proceso de validación técnica.

La validación analítica abarca la validación de datos, algoritmos y outputs. La Figura 3 recoge el objeto/alcance de cada fase del proceso de validación.

La validación clínica (figura 4) forma parte esencial del proceso de evaluación de un MDAI, y tiene lugar una vez que el sistema ha sido desplegado en su entorno previsto, ya sea real o simulado. Esta etapa comprende un procedimiento continuo para recopilar, valorar y analizar datos clínicos relacionados con el dispositivo, con el fin de determinar si existe evidencia clínica suficiente para confirmar el cumplimiento de los requisitos esenciales de seguridad y funcionamiento relativos a diseño y desarrollo recogidos en los reglamentos MDR y IVDR. Esta validación requiere no solo la evaluación estadística de resultados, sino también la demostración de utilidad clínica y de su relevancia para los flujos asistenciales. Para ello, diversas iniciativas metodológicas han desarrollado guías específicas para el diseño, implementación y evaluación de intervenciones basadas en IA en distintos tipos de estudio: STARD-AI para estudios diagnósticos, CONSORT-AI para ensayos clínicos aleatorizados, y SPIRIT-AI para protocolos de investigación.

Fase 2 Validación en entorno clínico	
Diseño del entorno simulado (Live in Lab)/real	**Actividades de validación**
El objetivo principal es evaluar el comportamiento del sistema en un contexto controlado pero representativo, sin comprometer la seguridad de pacientes reales.	Tiene como objetivo reproducir condiciones asistenciales reales en un contexto controlado, permitiendo evaluar el comportamiento del sistema de IA sin riesgo para pacientes reales. Las actividades están diseñadas para analizar la comprensibilidad, la usabilidad, el rendimiento clínico del producto y el papel del profesional sanitario como supervisor del sistema.

Figura 4. Validación clínica de PS con IA.

Las principales preocupaciones incluyen si el sistema ha sido validado en una población representativa del contexto clínico de uso (comparabilidad entre datos de entrenamiento y población objetivo), y si las salidas generadas tienen traducción clínica significativa, es decir, si aportan valor diagnóstico, pronóstico o terapéutico útil. En este sentido, resulta clave ir más allá de métricas clásicas como la exactitud, y evaluar aspectos como impacto clínico, aceptabilidad por parte del profesional sanitario, mejora de resultados en salud y aplicabilidad en condiciones del mundo real.

Parámetros para la validación

La validación de un producto sanitario que integra inteligencia artificial, especialmente si se basa en modelos fundacionales y está destinado a decisiones clínicas, debe apoyarse en una batería de parámetros cuantitativos y cualitativos que garanticen su aceptabilidad clínica, técnica y regulatoria. A continuación, se describen indicadores y criterios que se deben considerar.

1. Validación del rendimiento analítico: Esta evaluación tiene como objetivo garantizar que el dispositivo cumple con los requisitos de precisión, estabilidad y repetibilidad esperados en su aplicación clínica (Tabla 2).

Se busca responder a cuestiones fundamentales como:

- ¿El dispositivo proporciona resultados fiables y consistentes en diversos escenarios?
- ¿Mantiene su precisión incluso cuando hay cambios en los datos de entrada o en el entorno clínico?
- ¿Sus resultados están alineados con los estándares reconocidos en la práctica médica?

Además de las métricas cuantitativas específicas mostradas en la Tabla 2, es necesario considerar los diferentes niveles de validación implicados en cada componente del sistema. La Tabla 3 detalla estos parámetros, agrupados en tres bloques: validación de datos, validación de algoritmos y validación de outputs, señalando las herramientas y métricas aplicables en cada caso.

2. Evaluación de usabilidad: es recomendable seguir los principios establecidos en la norma ISO 62366, que proporciona un marco para aplicar principios de usabilidad en dispositivos médicos, con el objetivo de identificar, evaluar y reducir riesgos relacionados con el uso. Esta norma ayuda a asegurar que los aspectos de diseño centrados en el usuario minimicen fallos de manejo y favorezcan la seguridad, especialmente en dispositivos que incorporan inteligencia artificial, cuyos resultados pueden influir en decisiones clínicas críticas.

Métrica	Definición
ROC	Curva que muestra la relación entre la tasa de verdaderos positivos (sensibilidad) y la tasa de falsos positivos (1 - especificidad), permitiendo evaluar la capacidad de discriminación del modelo.
AUC	Área bajo la curva ROC
F1-score	F1 = 2 × (Precisión × Sensibilidad) / (Precisión + Sensibilidad)
Número de capas y neuronas	Determinan la complejidad y la capacidad de representación del modelo.
Learning rate/Tasa de aprendizaje	Es un valor numérico que refleja la magnitud con la que el modelo cambia arquitectura profunda para aprender mejor.

Tabla 2. Parámetros cuantitativos del rendimiento.

Parámetros a validar		
Validación datos	Representatividad de la población objetivo	Distribución de frecuencias por edad, sexo, raza, patología, comorbilidades
	Calidad del etiquetado.	Grado en que las etiquetas o anotaciones reflejan fielmente la realidad clínica. Indicadores: • Porcentaje de etiquetas verificadas por expertos. •Acuerdo interanotador (kappa > 0.80 deseable).
	Área bajo la curva ROC	patología
Validación de algoritmos	• Validación cruzada utilizando un conjunto de datos de entrenamiento y un conjunto de datos prueba de forma repetida, a fin de reducir la variabilidad y obtener una estimación más robusta del desempeño general del modelo (European Medicines Agency). • Métricas para la evaluación del rendimiento (AUC, F1, sensibilidad, etc), explicadas en la Tabla 2. • Técnicas de explicabilidad (LIME, SHAP) según el tipo de modelo.	
Validación outputs	Determinan la complejidad y la capacidad de representación del modelo.	
Learning rate/Tasa de aprendizaje	• Sensibilidad: TP / (TP + FN) • Especificad: TN / (TN + FP) • Precisión: TP / (TP + FP) • Tasa de Falsos Positivos: FP / (FP + TN) ó 1 - especificidad •Tasa de Falsos Negativos: FN / (FN + TP) ó 1 - sensibilidad	
TN: True Negatives (Verdaderos Negativos); TP: True Positives (Verdaderos Positivos); FP: False Positives (Falsos Positivos); FN: False Negatives (Falsos Negativos)		

Tabla 2. Parámetros cuantitativos del rendimiento.

Subgrupo evaluado	Objeto	Parametro de evaluación
Sexo y género	Detectar diferencias en el rendimiento del sistema entre hombres, mujeres	Sensibilidad, especificidad, PPV, F1-score por género
Edad	Comprobar si la edad influye en la precisión del sistema	Comparativa de métricas (ver apartado 5.1) entre grupos de edad
Condiciones clínicas coexistentes	Evaluar si las comorbilidades afectan la exactitud de las predicciones del modelo	Evaluación del rendimiento en presencia de múltiples factores clínicos (comorbilidades)
Origen étnico o cultural	Identificar posibles sesgos relacionados con poblaciones étnicas o culturales distintas	Análisis estratificado si se dispone de datos demográficos suficientes
Entorno asistencial	Validar la robustez del sistema en distintos niveles asistenciales como atención primaria, hospitalaria o remota	Análisis del rendimiento según tipo de centro o modalidad asistencial

Tabla 4. Parametros utilizados para evaluar las posibles diferencias entre subgrupos poblacionales.

En esta fase se examina la explicabilidad/comprensibilidad del producto. La norma ISO 62366 establece un marco para evaluar si:

• Los usuarios previstos comprenden e interactúan con el dispositivo de acuerdo con su diseño.

• La interfaz y la experiencia de usuario están optimizadas para facilitar su adopción en entornos clínicos.

3. Evaluación del rendimiento en subgrupos poblacionales: En el marco de la validación de sistemas de IA aplicados a la salud, resulta fundamental analizar el comportamiento del modelo en distintos subgrupos poblacionales, con el fin de garantizar el cumplimiento de los principios de equidad, no discriminación y robustez clínica. Esta necesidad está recogida en el AI Act, particularmente en los artículos 10 y 15, donde se establece que los sistemas de alto riesgo deben ofrecer garantías de desempeño uniforme en condiciones reales de uso y para todas las personas usuarias previstas. La tabla 4 recoge algunos parámetros habitualmente utilizados en esta evaluación.

Monitorización poscomercialización

El MDR y el AI Act exigen un sistema robusto de vigilancia poscomercialización y un plan de seguimiento clínico poscomercialización. Estos planes deben incluir:

- Monitorización continua del desempeño clínico.
- Detección de errores sistemáticos o efectos indeseados.
- Recolección de datos de uso real.

Conclusiones

El despliegue seguro y eficaz de APPs que incorporan LLM depende de cumplir estrictos requisitos regulatorios. Los marcos existentes, MDR en Europa, FDA en EEUU, proporcionan una base sólida, pero es imprescindible adaptarlos a las particularidades de los modelos fundacionales. Gestionar los cambios de forma planificada a través del plan de control de cambios predeterminado y establecer estrategias robustas de monitorización poscomercialización son pasos esenciales.

Los modelos fundacionales de IA abren una nueva etapa en los productos sanitarios, aportando innovación pero también importantes desafíos regulatorios. Su despliegue seguro exige una validación rigurosa de datos, algoritmos y outputs, garantizando su desempeño clínico conforme a marcos como MDR, AI Act y FDA. La validación, junto con una monitorización continua, es esencial para asegurar un uso responsable, eficaz y centrado en el paciente.

Los modelos fundacionales de IA abren una nueva etapa en los productos sanitarios, aportando innovación pero también importantes desafíos regulatorios. Su despliegue seguro exige una validación rigurosa de datos, algoritmos y outputs, garantizando su desempeño clínico conforme a marcos como MDR, AI Act y FDA.

Referencias

- 1. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, Arx S von, et al. On the Opportunities and Risks of Foundation Models [Internet]. arXiv; 2022 [citado 14 de julio de 2025]. Disponible en: http://arxiv.org/abs/2108.07258
- 2. Reglamento (UE) 2017/745 del Parlamento Europeo y del Consejo, de 5 de abril de 2017, sobre los productos sanitarios, por el que se modifican la Directiva 2001/83/CE, el Reglamento (CE) n.º 178/2002 y el Reglamento (CE) n.º 1223/2009 y por el que se derogan las Directivas 90/385/CEE y 93/42/CEE del Consejo. Diario Oficial de la Unión Europea, L 117, 5 de mayo,, 2017. p. 1–175.
- 3. Reglamento (UE) 2017/746 del Parlamento Europeo y del Consejo, de 5 de abril de 2017, sobre los productos sanitarios para diagnóstico in vitro y por el que se derogan la Directiva 98/79/CE y la Decisión 2010/227/UE de la Comisión. Diario Oficial de la Unión Europea, L 117, 5 de mayo, 2017. p. 176–332.
- 4. Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y se modifican determinados actos legislativos de la Unión. Diario Oficial de la Unión Europea, 2024/1689, 12 de junio, 2024.
- 5. U.S. Food and Drug Administration (FDA); Health Canada; United Kingdom Medicines and Healthcare products Regulatory Agency (MHRA). Good Machine Learning Practice for Medical Device Development: Guiding Principles. Silver Spring, MD: FDA; October 2021.
- 6. U.S. Food and Drug Administration (FDA). Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions: Guidance for Industry and Food and Drug Administration Staff. Silver Spring, MD: FDA; December 2024.
- 7. U.S. Food and Drug Administration (FDA). Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations. Draft Guidance for Industry and Food and Drug Administration Staff. Silver Spring, MD: FDA; January 7, 2025.
- 8. Saudi Food and Drug Authority (SFDA). Guidance on Artificial Intelligence (AI) and Machine Learning (ML) Technologies Based Medical Devices. MDS–G010, Version 1.0. Riyadh: SFDA; November 29 , 2022
- 9. Health Sciences Authority (HSA). Software as a Medical Device (SaMD). A Life Cycle Approach. Singapore: HSA; 2022.
- 10. Medical Device Coordination Group (MDCG). MDCG 2025-6: Interplay between the Medical Devices Regulation (MDR), the In Vitro Diagnostic Medical Devices Regulation (IVDR) and the Artificial Intelligence Act (AIA). Brussels: European Commission; june 2025.