CREDIBILIDAD DE LA IA FARMACÉUTICA: CLAVES DEL BORRADOR DE LA FDA

La FDA pone la credibilidad como uno de los ejes centrales para la aceptación de la inteligencia artificial (IA) en la industria farmacéutica. Este concepto, destacado en el borrador de la guía de la FDA, nos da las herramientas para documentar cómo los modelos de IA han sido desarrollados, entrenados y evaluados, ayudándonos así a determinar su riesgo de implementación y su adecuación al contexto de uso.

GUILLEM LÓPEZ,
CSV & QA Manager en OYTEC.

La Administración de Alimentos y Medicamentos de los Estados Unidos (FDA) emitió en enero de 2025 el borrador de la guía "Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products" que introduce un marco basado en riesgos para evaluar la credibilidad de los modelos de IA y ML.

Si bien la IA tiene el potencial de revolucionar el desarrollo, la evaluación y la producción farmacéutica al acelerar procesos y reducir costos, es imprescindible garantizar que los modelos que sustentan decisiones regulatorias sean fiables, transparentes y estén rigurosamente validados. Por ello, en este artículo exploraremos las consideraciones clave de la FDA sobre la credibilidad de los modelos de IA en el ciclo de vida farmacéutico, como calidad de datos, transparencia, validación y mantenimiento.

También destacaremos las estrategias prácticas para implementarlas en entornos regulados, alineando buenas prácticas (GxP), guías internacionales de referencia (p.ej.: GAMP5, Ed2; de la ISPE) y artículos académicos que abordan el tema.

¿Por qué necesitamos demostrar la credibilidad en los sistemas data-driven?

La introducción de modelos basados en datos, como la inteligencia artificial (IA) y el machine learning (ML), marca un cambio fundamental en el aseguramiento de la calidad. A diferencia de los sistemas tradicionales, diseñados con reglas predefinidas (código, scripts, triggers, ...), estos modelos aprenden patrones a partir de los datos históricos para generar predicciones, segmentaciones o clasificaciones entre otros. Este enfoque, aunque más potente y flexible, plantea un reto único: su desempeño ya no se basa en reglas claras, sino en patrones probabilísticos, lo que los hace más difíciles de auditar y validar.

Los modelos data-driven dependen completamente de la calidad de los datos utilizados para entrenarlos, lo que introduce riesgos significativos. Sesgos en los datos, falta de representatividad o cambios en las condiciones de operación (data drift) pueden comprometer su rendimiento. Además, su naturaleza probabilística genera resultados que varían dependiendo del contexto, complicando la reproducibilidad y la confianza en sus predicciones. Esto los diferencia radicalmente de los sistemas determinísticos, donde los resultados son siempre predecibles.

Por ello toma relevancia la **implementación de un Credibility Assessment Plan** cuyo objetivo es garantizar que los modelos de IA utilizados en entornos regulados sean **seguros, interpretables, confiables y alineados con el contexto de uso**. Para ello, el marco de credibilidad establecido en este borrador sienta en **siete pilares fundamentales**, proporcionando una metodología clara para el desarrollo del modelo a través de la descripción de los

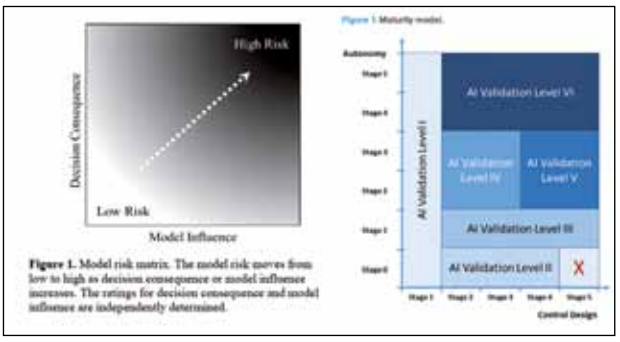

Diagrama 1. Clasificación del riesgo (FDA) y madurez del modelo (ISPE).

datos, la estructura, el entrenamiento, y la evaluación de la fiabilidad entre otros, garantizando que sus predicciones sean trazables y justificables dentro de su contexto de uso.

Marco de evaluación de credibilidad basado en el riesgo

El marco de **evaluación de credibilidad basado en el riesgo** descrito por la FDA es una metodología estructurada para garantizar que los modelos de IA empleados en decisiones regulatorias sean confiables, precisos y apropiados para su **contexto de uso (COU).** Este marco se centra en ajustar el nivel de rigor de la evaluación de credibilidad en función del riesgo inherente del modelo y del impacto de sus resultados en las decisiones regulatorias. El proceso consta de siete pasos clave:

1. Objetivo del modelo: Question of interest

La pregunta de interés define de manera precisa el problema, decisión o cuestión específica que el modelo de IA busca resolver. Responder a la pregunta de interés ayudará a guiar el diseño y desarrollo del modelo, asegurando que esté alineado claramente con su propósito.

Dicho de otra manera. Imagina que el modelo de IA es una herramienta. La pregunta de interés es la tarea para la que usarás esa herramienta.

¿Por qué es importante la "Question of interest"?

Una definición bien formulada nos permitirá delimitar el alcance del modelo, identificar las fuentes de datos necesarias y establecer las métricas que se utilizarán para evaluar su desempeño.

Además, esta pregunta conecta el modelo con su contexto de uso, ayudando a determinar cómo se integrará en los procesos existentes y qué decisiones dependen de sus resultados. También permitirá anticipar los requisitos regulatorios y las evidencias complementarias necesarias para validar el modelo.

2. Alcance y Business purpose: Context of use (COU)

El contexto de uso (COU) define el propósito y alcance del modelo de IA en un proceso específico. Describe qué será modelado, cómo se utilizarán los resultados y si el modelo funcionará junto a otras fuentes de evidencia.

El COU permite establecer si el modelo será la única herramienta de decisión o una complementaria, y cómo influirá en decisiones críticas del proceso. También delimita las condiciones bajo las cuales se utilizará, definiendo su integración en los procedimientos operativos, si las salidas serán verificadas externamente y su impacto en la calidad y seguridad.

El contexto de uso es el escenario donde actúa el modelo de IA. Delimita el papel del modelo, las reglas de la actuación y cómo interactúa con los demás elementos del entorno.

3. Riesgo del modelo

El riesgo del modelo se define como la posibilidad de que la salida del modelo lleve a una decisión incorrecta con consecuencias adversas. Este riesgo no está relacionado con el modelo en sí, sino con el impacto de sus resultados en el contexto de uso (COU).

El riesgo del modelo se define por dos factores clave: la **influencia del modelo y la consecuencia de la decisión.** La influencia refleja cuánto peso tiene la salida del modelo en el proceso de decisión, mientras que la consecuencia mide la gravedad de un resultado adverso si el modelo falla.

Buscando la equivalencia con la gestión de riesgos en los sistemas informatizados convencionales, podríamos entender la **Consecuencia** como la **Severidad**. Mientras que la **Influencia del modelo** podría ser equivalente a la **Detectabilidad**, entendiéndose como la disponibilidad de mecanismos que permitan contrastar si hay un fallo.

Determinar el riesgo del modelo permite adaptar las actividades de evaluación de credibilidad según su nivel de riesgo. Esto asegura que los recursos se enfoquen en los modelos con mayor impacto y necesidad de supervisión.

Llama la atención que la evaluación de riesgos propuesta por la FDA se limite a la Influencia y consecuencia, y no considere la evolución del proceso y sus datos (**Data drift**), y en consecuencia la pérdida de fiabilidad del modelo en el tiempo. Aunque este fenómeno si está contemplado durante la evaluación del modelo y el mantenimiento como veremos más adelante. En el artículo de la ISPE, "AI Maturity Model for GxP Application: A Foundation for AI Validation", se incluyen los conceptos de Diseño de controles que mide la independencia del modelo en la toma de decisiones y la Autonomía del modelo que sería la capacidad para auto actualizarse, que consideran la evolución del modelo en el tiempo. Esta disparidad en el

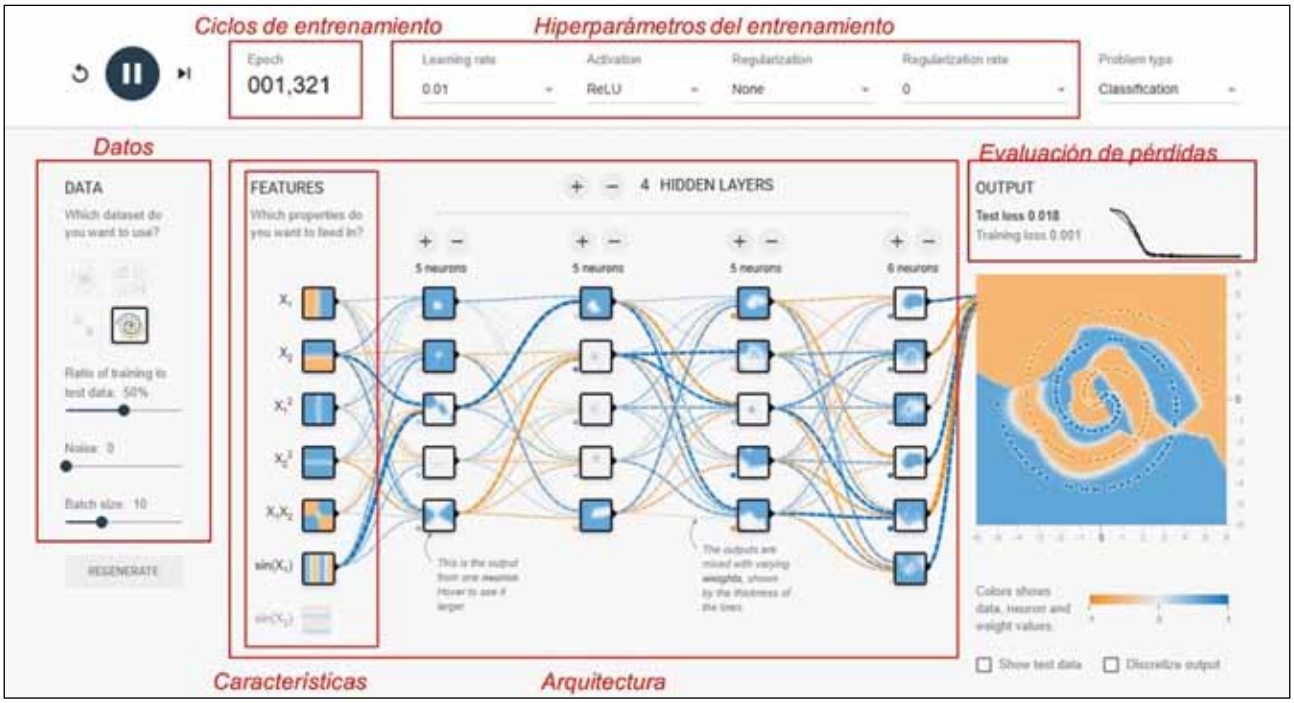

Diagrama 2. Ejemplo de entrenamiento de una red.

enfoque de evaluación nos da dos herramientas más que podrían usarse complementariamente. A continuación se muestran las matrices definidas por la FDA y el artículo de la ISPE.

Credibility assessment plan

El objetivo del plan es detallar cómo se establecerá la credibilidad de los resultados del modelo de IA dentro de su contexto de uso (COU). Se deberá ajustar según el nivel de riesgo del modelo y las actividades necesarias para garantizar su fiabilidad.

Debe incluir la información de los pasos previos: pregunta de interés, contexto de uso y evaluación del riesgo del modelo. Para modelos de bajo riesgo, el plan puede ser más simple; para modelos de alto riesgo, debe ser detallado y riguroso.

4.1 Descripción del modelo

El plan de evaluación de credibilidad debe incluir una descripción detallada del modelo de IA, comenzando por los **inputs y outputs** que utiliza y genera. Los inputs hacen referencia a los datos que alimentan el modelo, mientras que los outputs son los resultados específicos que se producen para responder a la pregunta de interés.

Además, es crucial detallar la **arquitectura del modelo**, como el uso de redes neuronales convolucionales u otras estructuras, así como los **parámetros internos** (hiperparámetros) que influyen en su funcionamiento, como pesos, tasas de aprendizaje o funciones de pérdida.

También se debe explicar el proceso de **selección de características** utilizadas en el modelo, asegurando que estas sean relevantes y representativas para el contexto de uso, así como las funciones de pérdidas usadas para medir el error en las predicciones y así poder optimizarlo.

Otro aspecto esencial es **justificar el enfoque metodológico** elegido para el desarrollo del modelo. Esto podría incluir explicar por qué se optó por una metodología específica (p.ej., supervisado, no supervisado) y concluye cómo esta se ajusta al problema que el modelo busca resolver.

Además, el plan debe incluir cualquier decisión técnica tomada durante el diseño, como la **elección de algoritmos y estrategias de optimización,** asegurando que estas decisiones estén alineadas con los requisitos del contexto regulatorio y operativo. Esta descripción detallada garantiza transparencia en el desarrollo del modelo, facilitando su validación y aceptación regulatoria.

4.2 Descripción de los Datos

Los datos de entrenamiento y ajuste son fundamentales para el desempeño y credibilidad de un modelo de IA. Estos datos incluyen información utilizada para enseñar al modelo cómo realizar predicciones y para optimizar sus parámetros internos. Por ello su origen, adquisición, procesado, y segregación en grupos es uno de los puntos diferenciales respecto a los modelos convencionales

Para la **Adquisición** de estos datos, se debe asegurar que sean representativos del problema que se busca resolver. Para que los datos sean relevantes y confiables,

es esencial que sean **precisos, completos y trazables.** Esto significa que los datos deben **reflejar las condiciones reales del contexto de uso,** contener la suficiente variedad para representar todos los escenarios posibles, y tener un historial claro que permita rastrear su origen y transformaciones.

Los datos deben ser **procesados** para garantizar su calidad, lo que incluye la eliminación de valores erróneos (limpieza) y la eliminación de duplicados entre otros, la normalización de escalas (transformación). Además, el anexo D11 de las GAMP5 incluye dentro del procesado de datos las acciones de: Perfilar (ej. formatear), Limpiar valores incorrectos, transformar (ej. homogeneizar unidades), Anonimizar (para GDPR [79]/privacidad de la UE) y aumentar para diversificar los datos.

Dentro del procesado de datos, el **etiquetado** es otro paso crítico, ya que define los grupos o resultados esperados que el modelo debe aprender a predecir, en especial en modelos con aprendizaje supervisado.

También debe indicarse si el **origen** de los datos es centralizado o si se utilizó un enfoque de *federated learning*, y detallar qué actividades de desarrollo del modelo se realizaron con cada conjunto de datos, así como la separación de los conjuntos de entrenamiento y evaluación.

En resumen, para garantizar que los datos de desarrollo sean adecuados para el **contexto de uso (COU),** es esencial que sean **relevantes y representativos** del entorno en el que se aplicará el modelo. Esto implica que los datos deben incluir elementos clave, contar con un número suficiente de muestras y reflejar con precisión el proceso de fabricación o el sistema que el modelo busca optimizar. Además, la **fiabilidad** de los datos debe asegurarse mediante su precisión, completitud y trazabilidad, permitiendo que cualquier transformación o uso en el desarrollo del modelo pueda ser verificado. También es importante documentar qué actividades específicas del desarrollo del modelo se realizaron con cada conjunto de datos, como entrenamiento, ajuste y validación, asegurando que su uso sea coherente con los objetivos del modelo y su contexto regulado.

4.3 Descripción del entrenamiento

El entrenamiento del modelo de IA y el ajuste de hiper parámetros es un proceso clave para su desarrollo y evaluación, en el que se ajustan sus parámetros para aprender patrones a partir de los datos. Existen diferentes metodologías de aprendizaje que pueden aplicarse según el tipo de problema.

En el aprendizaje supervisado, el modelo se entrena con datos etiquetados, lo que significa que conoce de antemano los resultados esperados y ajusta sus predicciones en consecuencia. En contraste, el aprendizaje no supervisado se emplea cuando no se cuenta con etiquetas y el modelo debe descubrir estructuras ocultas en los datos, como segmentar lotes de producción con características similares sin una clasificación predefinida. Aunque hay muchos otros tipos de entrenamiento.

Para evaluar el entrenamiento, se utilizan diversas métricas de desempeño que se detallan más adelante, las cuales permiten medir la precisión del modelo en su contexto de uso. Entre ellas, la curva ROC (Receiver Operating Characteristic) y el área bajo la curva (AUC) ayudan a evaluar el equilibrio entre verdaderos positivos y falsos positivos. Otras métricas, como la sensibilidad (capacidad del modelo para detectar correctamente los casos positivos) y la especificidad (habilidad para descartar correctamente los casos negativos), son esenciales en aplicaciones críticas, como la clasificación. Dependiendo del propósito del modelo, diferentes combinaciones de métricas pueden ser más relevantes para garantizar su desempeño óptimo en escenarios reales.

Uno de los desafíos principales en el entrenamiento del modelo es **evitar el infra/sobreajuste (under/overfitting), que ocurre cuando el modelo no generaliza bien (infra-juste)** o memoriza los datos de entrenamiento en lugar de generalizar patrones aplicables a nuevos datos. Para mitigar este problema, se aplican técnicas como la regularización, que introduce penalizaciones en el proceso de optimización para evitar que el modelo se ajuste excesivamente a los datos de entrenamiento.

En la descripción del entrenamiento deberemos reflejar también si se está empleando un **modelo pre-entrenado** justificando su adecuación, si el modelo trabaja solo o en conjunto (**"ensemble"**), si se han realizado técnica de calibración o ajuste (**"fine tuning"**), y las herramientas usadas para garantizar la calidad del entrenamiento (PNTs, control de versiones…).

La evaluación del modelo de IA nos permite garantizar que su desempeño sea adecuado dentro del contexto de uso (COU).

4.4 Descripción de la Evaluación

La evaluación del modelo de ML/IA deberá tener en consideración todos aquellos elementos que forman parte de él, como los algoritmos o los datos, pero también su desempeño en el contexto de uso, los usuarios que lo van a interactuar y las herramientas de aseguramiento de la calidad.

Evaluación de los datos:

Además, es importante documentar el proceso de recolección, procesamiento y anotación de estos datos, es especial si se ha llevado a cabo un aumentado de la muestra con datos sintéticos, así como la metodología utilizada para **mantener la independencia de los datasets** de entrenamiento y evaluación. En caso de solapamiento entre grupos de datos deberá justificarse su idoneidad.

Recordemos que los datos son el combustible del modelo, si no se realiza un proceso riguroso de

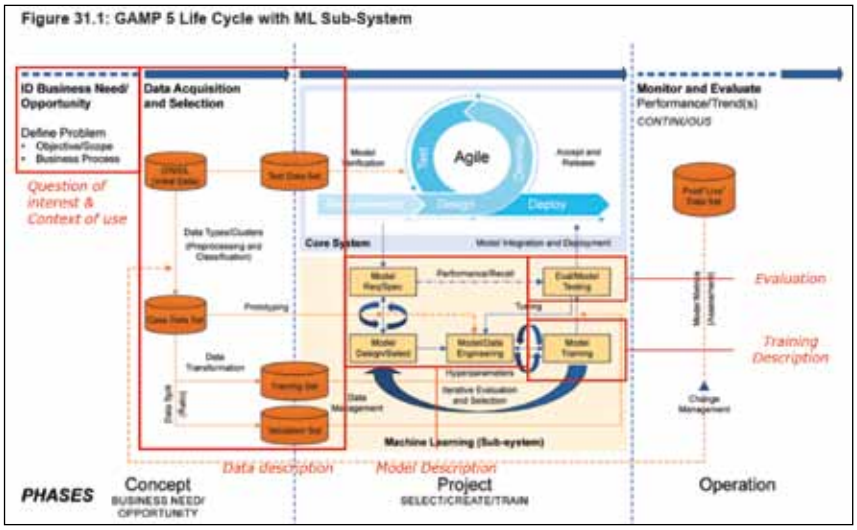

Diagrama 3. Ciclo de vida de un modelo de IA.

adquisición y procesado, podemos encontrarnos que el modelo no podrá generar bien los resultados en el entorno productivo.

Evaluación del Entorno

Un aspecto clave en esta evaluación es el fenómeno del **data drift**, que ocurre cuando los datos operativos difieren de los de entrenamiento, lo que puede afectar la precisión del modelo y requiere monitoreo continuo. Debemos asegurarnos de que los datos históricos se adecuen al contexto de uso del modelo.

Evaluación del modelo y las Métricas de evaluación

Para medir el rendimiento del modelo, se deben utilizar métricas estandarizadas que permitan evaluar su precisión y fiabilidad. La elección de las métricas deberá ser justificada.

Algunas métricas comunes de clasificación incluyen la **sensibilidad** (capacidad del modelo para detectar correctamente los casos positivos), la **especificidad** (capacidad de identificar correctamente los negativos), y la **curva ROC-AUC**, que evalúa el equilibrio entre tasas de verdaderos y falsos positivos.

También se debe analizar la **incertidumbre** en las predicciones, asegurando que el modelo proporcione estimaciones con un nivel de confianza adecuado. Si el modelo trabaja en conjunto con decisiones humanas (human-in-the-loop), la evaluación debe considerar la **interacción entre el modelo y los operadores**, midiendo el impacto de la IA en la toma de decisiones.

Identificación de sesgos

Además, es esencial identificar posibles **sesgos y limitaciones** del modelo antes de su implementación. Se deben evaluar patrones de error sistemáticos que puedan indicar algoritmic bias, por ejemplo, si el modelo favorece o discrimina a ciertos grupos debido a una representación inadecuada en los datos de entrenamiento.

A continuación, se muestra una comparativa gráfica entre el ciclo de vida del modelo de ML/AI propuesto en el Anexo D11 de las GAMP5 y los pasos descritos por el Credibility Assessment Plan.

4. Ejecución del Plan

Consiste en la implementación del plan de evaluación de credibilidad definido previamente. Antes de su ejecución, es recomendable discutirlo con la FDA para alinear expectativas y anticipar posibles desafíos. La ejecución debe ajustarse al riesgo del modelo y su contexto de uso, asegurando que cumpla con los requisitos regulatorios y operativos establecidos.

5. Documentar los Resultados y Desviaciones del Plan

Este paso se centra en registrar y formalizar los resultados de la evaluación de credibilidad del modelo de IA. Toda la información obtenida en los pasos 1 a 4 debe documentarse en un Credibility Assessment Report (CAR), que servirá como evidencia de que el modelo cumple con su contexto de uso (COU). Además, cualquier desviación del plan original debe ser identificada, justificada y explicada dentro del informe.

El CAR puede ser presentado a la FDA durante el periodo de adopción temprana como parte de una solicitud regulatoria, en un paquete de reuniones o mantenerse como documentación interna para inspecciones.

6. Determinar la Adecuación del Modelo para el Contexto de Uso

El último paso del marco de evaluación de credibilidad se centra en determinar si el modelo de IA es adecuado para su contexto de uso (COU).

Si el modelo cumple con los requisitos establecidos y demuestra un desempeño confiable dentro de su contexto, puede considerarse listo para su implementación en entornos regulados. Sin embargo, si la credibilidad no está suficientemente establecida, es necesario tomar medidas correctivas antes de su aprobación final.

Existen varias opciones para **mejorar la credibilidad** del modelo cuando no alcanza los estándares requeridos. Una estrategia es **reducir la influencia** del modelo,

complementando sus salidas con otras fuentes de evidencia para respaldar la toma de decisiones. También se puede **aumentar el rigor de la evaluación**, incorporando pruebas más estrictas o mejorando los procesos de evaluación. Otra alternativa es **ampliar los datos de desarrollo**, agregando información adicional para mejorar la representatividad del modelo. Además, pueden implementarse **controles adicionales** que mitiguen riesgos asociados a su desempeño, o incluso modificar el enfoque metodológico, ajustando la arquitectura o los algoritmos utilizados.

Si el modelo sigue sin cumplir con los requisitos de credibilidad, es posible que su contexto de uso deba ser revisado o modificado para garantizar que se aplique de manera segura y confiable. En los casos más extremos, esto podría significar la necesidad de redefinir su alcance o, incluso descartarse para entornos GxP.

Mantenimiento de la credibilidad tras la liberación

Al igual que en la mayoría de los elementos GMP, la calidad debe asegurarse a lo largo de todo el ciclo de vida. Para ello, se somete a revisiones periódicas y a planes de acción que garanticen su mantenimiento durante toda su vida útil.

En el caso de los modelos de IA, este mantenimiento de la credibilidad—ya sea a través de revisiones periódicas, verificaciones continuas o cualquier otro enfoque—cobra especial importancia debido al fenómeno de *data drift* y a la evolución de los procesos y sus datos. Es decir, los datos históricos utilizados para crear, entrenar y evaluar el modelo podrían no ser representativos del proceso actual. Como consecuencia, si no se realizan acciones de reentrenamiento y reevaluación con datos adecuados, se podría perder parte de la fiabilidad del modelo.

Las actualizaciones de los modelos de IA pueden darse de forma deliberada o incluso automática, ya que algunas plataformas son capaces de evolucionar sin intervención humana. Por ello, las empresas deben anticiparse y prepararse para modificaciones que puedan presentarse tanto por la dinámica interna del modelo como por ajustes necesarios en los procesos de fabricación o distribución. Este proceso requiere planes de acción claros que incluyan indicadores de rendimiento, protocolos de validación y estrategias de comunicación para informar a las autoridades regulatorias sobre cualquier modificación relevante, este proceso puede basarse en la guía ICH Q12.

Comunicación con la FDA

La FDA recomienda encarecidamente que los patrocinadores y demás partes interesadas establezcan un compromiso temprano con la Agencia para definir las actividades de evaluación de credibilidad adecuadas al nivel de riesgo y al contexto de uso del modelo de IA, así como para anticipar y resolver posibles complicaciones. Para ello, ofrece opciones como reuniones con los centros responsables (p. ej., INTERACT, CBER/CDER, Pre-IND).

Referencias

• Food and Drug Administration. (2025). Draft. Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products. U.S. Department of Health and Human Services. https://www.fda.gov/media/184830/download

• International Society for Pharmaceutical Engineering. (2022). GAMP 5 Guide: A Risk-Based Approach to Compliant GxP Computerized Systems (2nd ed.). Appendix D11.

• Erdmann, N., Blumenthal, R., Baumann, I., & Kaufmann, M. (2022). AI Maturity Model for GxP Application: A Foundation for AI Validation. Pharmaceutical Engineering, March/April 2022. International Society for Pharmaceutical Engineering. https://ispe.org/pharmaceutical-engineering/march-april-2022/ai-maturity-model-gxp-application-foundation-ai

• Food and Drug Administration. (2022). Draft Guidance for Industry and Food and Drug Administration Staff: Computer Software Assurance for Production and Quality System Software. U.S. Department of Health and Human Services. https://www.fda.gov/media/161521/download

• Food and Drug Administration. (2023). Assessing the Credibility of Computational Modeling and Simulation in Medical Device Submissions: Guidance for Industry and Food and Drug Administration Staff. U.S. Department of Health and Human Services. https://www.fda.gov/media/154985/download lizado.

Glosario

• CAP - Plan de aseguramiento de la credibilidad (del inglés, Credibility Assessment Plan)

• CAR - Informe de aseguramiento de la credibilidad (del inglés, Credibility Assessment Report)

• COU - Contexto de uso (del inglés, Context Of Use)

• FDA - Administración de medicamentos y alimentos (del inglés, Food and Drug Administration)

• GAMP5 - Guía de buenas prácticas de automatización

• GMP - Normas de correcta fabricación (del inglés, Good Manufacturing Practices)

• GxP - Buenas prácticas (del inglés, Good "x" Practices)

• IA - Inteligencia artificial

• ICH- Consejo internacional para la armonización

• ISPE - Sociedad internacional de ingeniería farmacéutica

• ML - Aprendizaje automático (del inglés, Machine Learning)

• RA - Análisis de Riesgos