AEQUITAS

D2.1 Informe de sesgos

Cofinanciado por
la Unión Europea

Índice

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanciado por la Unión Europea

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanciado por la Unión Europea

AEQUITAS

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Introducción

Parte del proyecto AEQUITAS consiste en crear una base de datos sobre los sesgos de género y raciales en las aplicaciones médicas de la inteligencia artificial (IA), centrándose específicamente en tres enfermedades: las enfermedades cardiovasculares, la diabetes y la depresión.

Para completar esta tarea, las entidades socias del consorcio han tenido que recopilar primero diversas fuentes sobre los sesgos mencionados anteriormente. El Hospital Universitario de Colonia (Universitätsklinikum Köln, UKK), como líder de la tarea y experto en la materia, organizó la actividad de recopilación de información y proporcionó la plantilla de mapeo, que los socios y socias utilizaron para mapear las fuentes, asegurando que la información relevante pudiera transferirse fácilmente a la base de datos.

Este informe presenta los fundamentos teóricos y científicos que guiaron la selección de la actividad de recopilación y la plantilla de mapeo, acompañados de estudios de casos que muestran los diferentes tipos de sesgos; las implicaciones políticas de que los sesgos inducidos por la IA biomédica afecten a los derechos protegidos por la Carta de los Derechos Fundamentales de la Unión Europea; una descripción de la actividad de recopilación de datos; la plantilla de mapeo; una lista de fuentes recopiladas por los socios de AEQUITAS; y otro material de apoyo. El resto del informe se estructura de la siguiente manera:

En primer lugar, presentamos los fundamentos teóricos de nuestro trabajo en una introducción sobre las aplicaciones médicas de la IA y el concepto de sesgo en los sistemas informáticos y la medicina. Comenzamos centrándonos en la medicina, presentando en primer lugar cómo se manifiestan los sesgos raciales y de género en la atención médica y, en segundo lugar, cómo la bioética aborda las cuestiones morales que surgen en la práctica de la medicina y la investigación biomédica, ofreciendo una

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

breve introducción a los cuatro principios de la bioética (autonomía, no maleficencia, beneficencia y justicia).

A continuación, pasamos al ámbito de la IA, presentando los tipos de sesgo que se pueden observar en los sistemas de IA tal y como se manifiestan en el proceso de aprendizaje automático e inteligencia artificial (ML/AI). Cada tipo de sesgo va acompañado de ejemplos y un estudio de caso sobre los sesgos de género y raza y su impacto a nivel social en relación con las tres enfermedades en las que se centra el proyecto AEQUITAS (enfermedades cardiovasculares, diabetes y depresión), extraídos de fuentes recopiladas por los socios de AEQUITAS tras la finalización de la tarea T2.2. Cuando esto no fue posible porque el material recopilado no demostraba claramente el tipo específico de sesgo de IA considerado, se presentó un caso alternativo de otro ámbito médico que se podía generalizar fácilmente a las enfermedades objetivo de AEQUITAS. Las descripciones de los estudios de casos, junto con las fuentes recopiladas, se basan en recursos científicos adicionales según han sido necesario para respaldarlas.

Por último, en la sección «Implicaciones políticas», se muestran cómo los distintos tipos de sesgos de la IA afectan a los derechos fundamentales protegidos por la Carta de la UE, en particular los preceptos de dignidad humana, igualdad ante la ley, no discriminación, así como el derecho a la integridad de la persona, el derecho a la asistencia sanitaria, la protección de datos y el derecho a un recurso efectivo, concluyendo con las salvaguardias que pueden establecerse en las evaluaciones de conformidad, la supervisión posterior a la comercialización y la contratación pública.

El informe concluye con las referencias y los siguientes apéndices:

Apéndice 1: Recopilación de fuentes y método de mapeo, que contiene la plantilla de mapeo y describe el proceso de recopilación, mapeo y evaluación de la información llevado a cabo durante las tareas T2.1 y T2.2.

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Anexo 2: Contiene material de apoyo para las tareas T2.1 y T2.2, es decir, diapositivas de las reuniones de los socios en las que se describe el proceso, presentadas por UKK.

Anexo 3: Lista de fuentes recopiladas por los socios de AEQUITAS.

Aplicaciones médicas de la IA

El auge de las aplicaciones de IA en los últimos años ha tenido un gran impacto en la medicina, incluyendo la adquisición de datos digitalizados, el aprendizaje automático y la infraestructura informática (Yu et al., 2018). En particular, la introducción de algoritmos de aprendizaje profundo en áreas como la visión artificial y el procesamiento del lenguaje natural ha revolucionado las aplicaciones informáticas en radiología, patología, cardiología, diabetología, psiquiatría, oncología, etc. (Esteva et al., 2019; Koteluk et al., 2021; Rajpurkar et al., 2022; Gou et al., 2024). La Organización Mundial de la Salud (OMS) enumera los siguientes ámbitos de aplicación de los sistemas de IA en la atención sanitaria: diagnóstico y diagnóstico basado en predicciones, atención clínica, investigación y desarrollo de fármacos, gestión y planificación de sistemas sanitarios, salud pública y vigilancia de la salud pública, promoción de la salud, prevención de enfermedades, vigilancia basada en predicciones, preparación para emergencias y respuesta a brotes epidémicos (Organización Mundial de la Salud, 2021).

Sin embargo, la llegada de las aplicaciones de IA a la medicina conlleva una serie de retos, como los retos de implementación, entre los que se incluyen la confianza en los modelos y las limitaciones de los datos, las cuestiones de responsabilidad, que incluyen los retos normativos y la atribución adecuada de responsabilidades, y la garantía de la equidad mediante el uso ético de los datos, la distribución equitativa de los beneficios y la detección y mitigación de los sesgos (Rajpurkar et al., 2022).

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanciado por la Unión Europea

AEQUITAS

El proyecto AEQUITAS se centra en los casos de sesgo de género y raza en las enfermedades cardiovasculares, la diabetes y la depresión. Las aplicaciones médicas de IA apoyan la atención cardiovascular mediante el apoyo a la toma de decisiones clínicas, la telemedicina, la evaluación de riesgos, la terapia personalizada, el análisis predictivo y la monitorización remota (Bernstein et al., 2025; Naskar et al., 2025), mejoran el control de la diabetes (incluida la monitorización de los pacientes y el autocontrol), el diagnóstico, el tratamiento y la prevención (Contreras y Vehi, 2018; Khalifa y Albadawy, 2024; Naskar et al., 2025; Sheng et al., 2024). En cuanto a la depresión, participan en la detección, el diagnóstico y el tratamiento (Alhuwaydi, 2024), con especial atención a la detección y el cribado mediante el uso de modelos de lenguaje grandes (LLM) (Cao et al., 2025; Kumari et al., 2025; Mao et al., 2023; Wang et al., 2025). En todas las áreas mencionadas, existen retos relacionados con los sesgos, por ejemplo, en lo que respecta a las enfermedades cardiovasculares, la diabetes y la depresión, véase van Assen et al. (2024), Cronjé et al. (2023) y Dang et al. (2024), respectivamente.

Los retos como el sesgo, ya sea en medicina o en IA, se abordan mediante una combinación de bioética y ética de la IA. En la siguiente sección, ofrecemos una breve descripción general de estos dos tipos de dominios de ética aplicada que sirvieron de base teórica y científica para desarrollar la plantilla de mapeo de sesgos.

Ética y sesgos en la medicina y la IA

Bioética y sesgos en la medicina

Los sesgos en la medicina están bien documentados: véase, por ejemplo, Hammond et al. (2021) en relación con los sesgos cognitivos, que consisten en errores sistemáticos en el pensamiento debidos a limitaciones del procesamiento humano o a patrones mentales inadecuados, y FitzGerald y Hurst (2017) en relación con los sesgos implícitos que implican asociaciones ajenas a la conciencia y que conducen a una evaluación

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

negativa de una persona sobre la base de características irrelevantes, como la raza o el género.

El sesgo racial en la medicina está bien estudiado en el caso de los Estados Unidos, por ejemplo, donde se ha documentado que las personas afroamericanas, así como los miembros de otros grupos minoritarios, son sometidas menos procedimientos y reciben una atención médica de peor calidad, al obtener un tratamiento menos agresivo, experimentar tasas más bajas de tratamiento quirúrgico y recibir menos derivaciones a especialistas que las personas blancas (Bowser, 2001; Williams y Wyatt, 2015).

El sesgo de género puede atribuirse a la ceguera de género y a los prejuicios estereotipados sobre hombres y mujeres (Hamberg, 2008), a lo que se suma un desconocimiento generalizado sobre el funcionamiento del cuerpo femenino y sus diferencias biológicas con respecto al cuerpo masculino. Por ejemplo, se ha visto que las mujeres de 50 años o más en estado crítico tenían menos probabilidades que los hombres en estado crítico de ser ingresadas en una unidad de cuidados intensivos (UCI) (Bierman, 2007), e incluso los modelos de ratones machos están, en general, más representados que los modelos hembras en la investigación biomédica básica, preclínica y quirúrgica (Yoon et al., 2014).

También es importante señalar que las personas LGBT+ sufren discriminación en el acceso a la atención sanitaria y son objeto de estereotipos que no afectan a la población heterosexual. Estos factores sociales y culturales perpetúan la discriminación y repercuten en la salud. Por ejemplo, un estudio realizado en los Estados Unidos, basado en datos de la Encuesta Nacional de Salud (NHIS) de 2013-2014, reveló que los adultos LGB declaraban niveles más altos de mala salud, limitaciones funcionales, angustia psicológica grave y dificultades para pagar la atención sanitaria en comparación con sus homólogos heterosexuales. Estas

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

desigualdades se deben al estrés de los grupos minoritarios y a la marginación social multifacética (Liu et al., 2023).

Por otra parte, la medicina como disciplina está sujeta a altos estándares éticos desde la antigüedad hasta la actualidad (Baker y McCullough, 2008). Durante siglos, ha existido la expectativa social de que los médicos sigan las normas éticas de responsabilidad profesional establecidas por los estándares de su profesión, tal y como se manifiesta a través de normas profesionales que van desde el juramento hipocrático del 400 a. C. (Miles, 2005) hasta las Declaraciones de Ginebra y Helsinki (Tröhler, 2008). Como señalan (Vevaina et al., 1993), los médicos son responsables de cumplir el código ético de su profesión debido a la inversión que la sociedad realiza en su formación (económica y el uso de sus miembros como material didáctico a lo largo de la formación y la carrera del médico) y al monopolio virtual que se concede a su profesión a través de la concesión de licencias.

La ética biomédica (o bioética) es un ámbito de la ética práctica (o aplicada) que aborda las cuestiones morales que surgen en la práctica de la medicina y la investigación biomédica (Vevaina et al., 1993). En el centro de la ética biomédica se encuentran los cuatro principios definidos por Beauchamp y Childress (Beauchamp y Childress, 2019):

1. Autonomía: respetar la capacidad de toma de decisiones de las personas autónomas. Hay dos condiciones generales esenciales para la autonomía: la libertad, que se manifiesta como independencia de influencias controladoras, y la agencia, es decir, la capacidad de actuar de forma intencionada.
2. No maleficencia: evitar causar daño.
3. Beneficencia: tomar medidas positivas para ayudar a los demás, concretamente, prevenir el mal o el daño, eliminar el mal o el daño y promover el bien.

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

4. Justicia: distribuir los beneficios, los riesgos y los costes de manera equitativa. La justicia se interpreta como un trato justo, equitativo y adecuado para las personas y los grupos, dadas las numerosas disparidades en la atención sanitaria y la investigación basadas en la raza, el origen étnico, el género y la condición social.

Ética y sesgos de la IA

La introducción de la IA y el rápido desarrollo de sus aplicaciones han planteado diversas cuestiones éticas (Christoforaki y Beyan, 2022), entre las que destacan los sesgos y la discriminación.

Así, la ética de la IA se desarrolló como un ámbito de la ética práctica (o aplicada) que comprende «un conjunto de valores, principios y técnicas que emplean normas ampliamente aceptadas de lo que está bien y lo que está mal para guiar la conducta moral en el desarrollo y el uso de las tecnologías de IA» (Leslie, 2019, p. 3).

La ética de la IA se basa tanto en la bioética (los cuatro principios presentados anteriormente) como en el discurso de los derechos humanos, que incluye, entre otras cosas, el derecho a la igualdad, libertad y dignidad ante la ley, la protección de los derechos civiles, políticos y sociales, el reconocimiento universal de la personalidad y el derecho a participar libremente y sin trabas en la vida de la comunidad (Leslie, 2019).

Los cuatro principios de bioética modificados con la explicabilidad se traducen para la IA en (Floridi et al., 2018) de la siguiente manera:

1. Autonomía, como el poder de los seres humanos para decidir o no decidir, y que conlleva el riesgo de delegar demasiado en las máquinas.
2. No maleficencia, como la prevención de daños derivados de la intención de los seres humanos o del comportamiento impredecible de las máquinas.

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

3. Beneficencia, como la promoción del bienestar, la preservación de la dignidad y la sostenibilidad del planeta.
4. Justicia, como la prevención y eliminación de las discriminaciones injustas ya existentes, así como de nuevos daños, y la garantía de la distribución equitativa de los beneficios de la IA.
5. Explicabilidad, definida como la comprensión y la rendición de cuentas de los procesos de toma de decisiones de la IA.

En lo que respecta a los derechos humanos, según un informe de 2018 financiado por el Consejo de Europa (Comité de expertos en intermediarios de Internet (MSI-NET), 2018), los derechos humanos que se ven especialmente afectados por los algoritmos y las técnicas de tratamiento automatizado de datos son los siguientes:

- Juicio libre y debido proceso
- Privacidad y protección de datos
- Libertad de expresión
- Recurso efectivo
- Libertad de reunión y asociación
- Prohibición de la discriminación
- Derechos sociales y acceso a los servicios públicos
- Derecho a elecciones libres

Los algoritmos sesgados se mencionan explícitamente como posibles factores de discriminación contra grupos sociales por motivos de edad, orientación sexual, raza, género o situación socioeconómica (Comité de expertos en intermediarios de Internet (MSI-NET), 2018, p. 27). Además, el Convenio Marco del Consejo de Europa sobre Inteligencia Artificial y Derechos Humanos, Democracia y Estado de Derecho menciona específicamente que los Estados miembros «adoptarán o mantendrán medidas con miras a garantizar que las actividades dentro del ciclo de vida de los sistemas de inteligencia artificial respeten la igualdad, incluida la igualdad de género, y la

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

prohibición de la discriminación, según lo dispuesto en el derecho internacional y nacional aplicable», (Convenio Marco del Consejo de Europa sobre Inteligencia Artificial y Derechos Humanos, Democracia y Estado de Derecho, 2024, p. 4).

Así, la ética de la IA también ha convergido en un conjunto de principios basados en los cuatro principios clásicos de la ética médica, así como en otros enfoques, resumidos en (Christoforaki y Beyan, 2022). Sin embargo, como se señala en (Mittelstadt, 2019), en comparación con la medicina, el desarrollo de la IA carece de: (1) objetivos comunes y obligaciones fiduciarias, (2) historial y normas profesionales, (3) métodos probados para traducir los principios en la práctica, y (4) mecanismos sólidos de responsabilidad jurídica y profesional; esto socava el éxito del enfoque basado en principios.

Naturalmente, también existe un complejo panorama normativo que regula el desarrollo y el uso de la IA en la UE, incluidas las leyes contra la discriminación, un tema que, sin embargo, queda fuera del alcance de este informe.

Las organizaciones de la sociedad civil (OSC), como partes interesadas en el ecosistema de la atención sanitaria (Vayena et al., 2018), pueden desempeñar un papel importante en la identificación y el tratamiento de los sesgos de la IA y la gobernanza de la IA en general, mediante la promoción del desarrollo ético de la IA, la rendición de cuentas de las partes interesadas, la educación del público, la representación de las comunidades marginadas, la configuración de marcos normativos y reglamentarios, y el fomento de la colaboración entre los gobiernos, las empresas tecnológicas y el público (Korir, 2024).

Dentro de este marco teórico, se han desarrollado explícitamente diversas soluciones técnicas para abordar los sesgos. En la siguiente sección, presentamos una clasificación de los sesgos inducidos por la IA que sirvió de base para nuestra plantilla de mapeo, centrándonos en su impacto en la discriminación de género y racial. Los sesgos mentales humanos (Hofmann, 2023), por ejemplo, los sesgos cognitivos, como el sesgo

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

de confirmación o el sesgo de disponibilidad, aunque tienen un gran impacto en la medicina, se consideran fuera del alcance del presente proyecto.

Sesgo en los sistemas de IA

El sesgo en los sistemas informáticos se define en (Friedman y Nissenbaum, 1996, p. 332) como un término «[que se refiere] a los sistemas informáticos que discriminan de forma sistemática e injusta a determinadas personas o grupos de personas en favor de otros. Un sistema discrimina injustamente si niega una oportunidad o un bien, o si asigna un resultado indeseable a un individuo o grupo de individuos por motivos irrazonables o inapropiados».

Según (Friedman y Nissenbaum, 1996), el sesgo en los sistemas informáticos se puede distinguir en tres categorías: sesgo preexistente, sesgo técnico y sesgo emergente. En la siguiente subsección, examinamos cada tipo de sesgo y lo ilustramos con estudios de casos, tal y como se manifiesta en la literatura científica.

Sesgo preexistente

El sesgo preexistente proviene de sesgos en las instituciones sociales, las prácticas y las actitudes que ya existen y son independientes y, por lo general, están presentes antes de la creación del sistema. Este tipo de sesgo se incorpora al sistema de forma consciente o inconsciente, a veces incluso cuando los creadores del sistema intentan evitarlo.

Estudio de caso: Diagnóstico de enfermedades cardiovasculares en mujeres

Las enfermedades cardiovasculares (ECV) se han percibido comúnmente como «enfermedades masculinas», lo que ha contribuido a un infradiagnóstico y un tratamiento insuficiente en las mujeres. Como se muestra en (Al Hamid et al., 2024), una revisión sistemática sobre el tema, las ECV se notificaban menos entre las mujeres

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

que presentaban síntomas más leves que los hombres o cuyos síntomas se diagnosticaban erróneamente como síntomas gastrointestinales o relacionados con la ansiedad; por lo tanto, a las mujeres se les ofrecían menos pruebas diagnósticas y medicamentos, y se las derivaba a cardiólogos o se las hospitalizaba con menos frecuencia. Además, en caso de hospitalización, las mujeres tenían menos probabilidades de recibir una intervención coronaria. Por lo tanto, los médicos, especialmente los hombres, subestimaban los factores de riesgo de las mujeres. Dado que las mujeres siguen estando infrarrepresentadas en el campo de la cardiología (Fatunde et al., 2025), se puede concluir que las mujeres tienen menos probabilidades de recibir una atención sanitaria adecuada debido a los prejuicios ya existentes.

Los sistemas de IA se entrenan utilizando datos recopilados por las prácticas existentes, por lo que un sistema de diagnóstico de ECV basado en IA incorporará este sesgo, creando discriminación contra las mujeres, independientemente de las decisiones que se tomen durante la implementación técnica.

Sesgo técnico

El sesgo técnico surge de limitaciones o consideraciones técnicas, especialmente cuando los creadores de sistemas intentan adaptar las constructos humanos a los ordenadores, como cuantificar lo cualitativo, discrecionar lo continuo o formalizar lo informal. Además, descontextualizar los algoritmos de los entornos en los que operan puede hacer que no traten a todos los grupos de forma justa en todas las circunstancias importantes.

Estudio de caso: Precisión predictiva de los modelos de predicción del riesgo de accidente cerebrovascular en poblaciones negras y blancas

(Hong et al., 2023) realizaron un estudio retrospectivo sobre la precisión predictiva del riesgo de accidente cerebrovascular comparando los modelos de predicción de riesgo específicos para accidentes cerebrovasculares existentes y las nuevas técnicas de

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

aprendizaje automático que incluyen, entre otros criterios, la raza de los pacientes. Todos los algoritmos mostraron una peor capacidad de diferenciación en las personas de raza negra que en las de raza blanca. Según los autores, esta situación puede atribuirse a factores de riesgo no recogidos en los datos, como el tipo de seguro, las barreras lingüísticas y otros factores derivados del acceso diferencial a los servicios de atención sanitaria, es decir, los datos están descontextualizados del entorno socioeconómico en el que se produjeron. Al mismo tiempo, todos los factores de riesgo mencionados son constructos difíciles de representar de forma que puedan ser procesados por ordenadores. A todo lo anterior, podríamos añadir que los algoritmos de IA de última generación son, por naturaleza, opacos en cuanto a las características que seleccionan para lograr una alta precisión (Knight, 2017), lo que hace que incluso sus creadores sean incapaces de explicar cómo funcionan y, por lo tanto, de controlar si alguno de los factores socioeconómicos mencionados anteriormente se tiene realmente en cuenta en el funcionamiento interno del sistema de IA.

Sesgo emergente

El sesgo emergente se manifiesta en un contexto de uso con usuarios reales, normalmente después de que se haya completado el diseño, como resultado de cambios en el conocimiento social que no pueden incorporarse, o no se incorporan, al diseño del sistema, o de una población con conocimientos o valores culturales diferentes a los asumidos en el diseño.

Estudio de caso: Cambios en los conjuntos de datos

Un cambio en el conjunto de datos es una discrepancia entre las distribuciones de los conjuntos de datos de entrenamiento y prueba durante el desarrollo del algoritmo y puede dar lugar a un rendimiento desigual a nivel de subgrupos (Chen et al., 2023).

En la detección del cáncer de piel, por ejemplo, muchos conjuntos de datos de imágenes utilizados para entrenar algoritmos de IA para detectar el cáncer de piel

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

proceden de países con poblaciones de piel clara (Guo et al., 2021), lo que supone una representación insuficiente de determinados grupos demográficos. Los algoritmos de IA entrenados con estos conjuntos de datos tienen un rendimiento inferior cuando se aplican en países con una población más diversa, lo que discrimina a las personas de piel oscura. Los conjuntos de datos son difíciles y costosos de recopilar, anotar y validar, lo que hace necesario que los sistemas de IA desarrollados en países de ingresos bajos y medios dependan de conjuntos de datos disponibles públicamente que pueden no reflejar la distribución de su población, lo que da lugar a un desajuste entre las poblaciones de origen y de destino. Lo mismo puede ocurrir también en países de ingresos altos, por ejemplo, debido a los cambios demográficos provocados por el aumento de la inmigración o a las variaciones en la autoidentificación racial. Como se señala en (Chen et al., 2023), «dado que ahora se acepta que la raza es una construcción social y que existe una mayor variabilidad genética dentro de una raza concreta que entre razas» […] «la comunidad médica ha comenzado a darse cuenta de que las taxonomías del pasado no representan adecuadamente a los grupos de personas a los que pretenden referirse» y «pueden ocultar la cultura, la historia, el estatus socioeconómico y otros factores que confunden la equidad».

Tipos de sesgos específicos del proceso de ML/AL

Si bien lo anterior es válido para todos los sistemas informáticos, las aplicaciones de IA tienen requisitos más específicos, por lo que necesitábamos una taxonomía más detallada. En consecuencia, decidimos seguir la clasificación de sesgos presentada en (Suresh y Guttag, 2020), ya que identifica los tipos de sesgos en cada paso del proceso de ML/IA, como se ilustra en la figura 1.

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

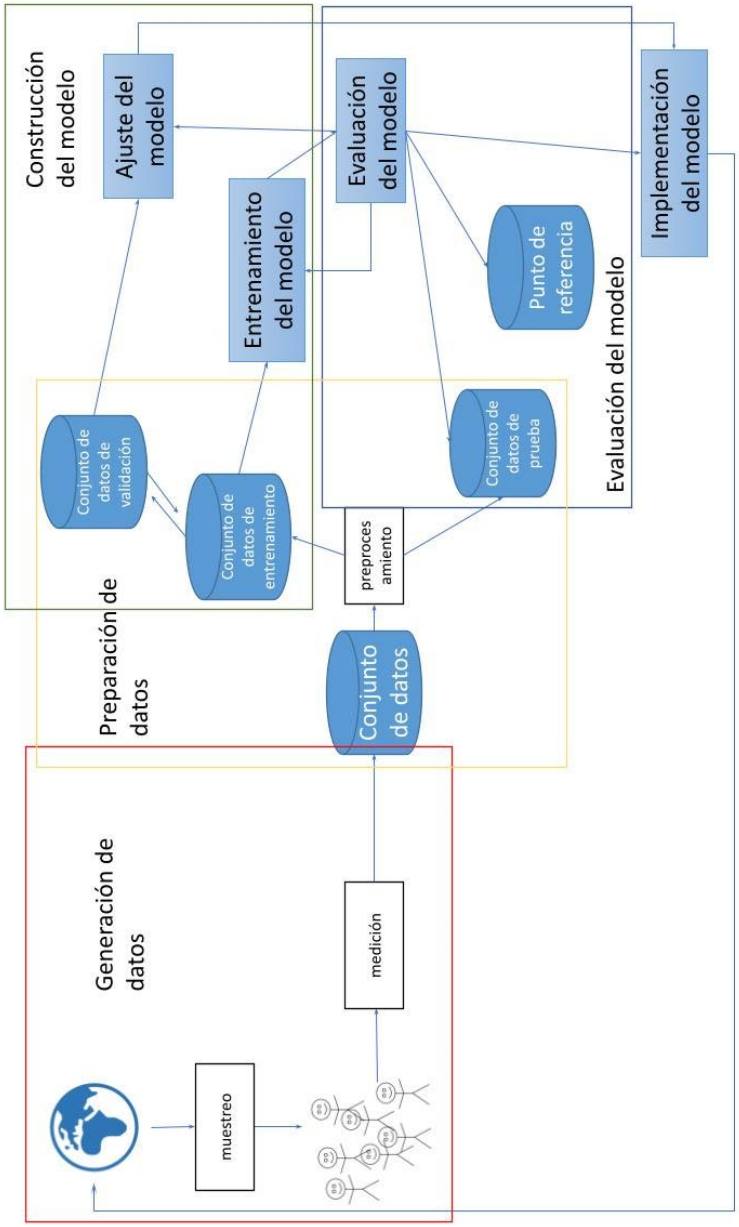

Imagen 1. Proceso de ML/IA. Imagen adaptada de (Suresh y Guttag, 2020).

Un proceso típico de ML/IA se puede describir de la siguiente manera:

- **Generación de datos.** La creación de un sistema de ML/IA comienza con la generación de datos. Esto implica, en primer lugar, recopilar y preparar datos

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

para recopilar un conjunto de datos para el sistema de IA. Los datos existentes en el mundo deben recopilarse identificando una muestra de la población objetivo. El siguiente paso es definir y medir las características relevantes para la aplicación que se va a implementar y/o anotar los datos con las etiquetas adecuadas. Se trata de un proceso costoso y largo, por lo que, en la mayoría de los casos, los profesionales de la IA utilizan conjuntos de datos existentes (ya sean públicos o comprados)

- **Preparación de datos.** En esta etapa, el conjunto de datos se divide en tres partes: el *conjunto de datos de entrenamiento*, que es el conjunto de datos real utilizado para entrenar el modelo; el *conjunto de datos de validación*, que es una muestra de datos utilizada para evaluar el ajuste del modelo al conjunto de datos de entrenamiento mientras se ajustan los hiperparámetros del modelo (parámetros del modelo que no se pueden aprender a partir de los datos, por ejemplo, el número de capas y neuronas en un modelo de red neuronal). En esta fase, es posible que sea necesario preprocesar los datos (por ejemplo, limpiarlos o normalizarlos); y el *conjunto de datos de prueba*, la parte de los datos que se utiliza para evaluar el modelo final y que proporciona un estándar de referencia una vez que el modelo está completamente entrenado.
- **Creación del modelo.** En esta fase, el modelo se entrena con los datos de entrenamiento y se ajusta mediante la modificación de los hiperparámetros en el conjunto de datos de validación.
- **Evaluación del modelo.** El modelo entrenado se evalúa utilizando el conjunto de datos de prueba y, en ocasiones, conjuntos de datos de referencia, que son conjuntos de datos compilados de forma independiente que se utilizan para demostrar la solidez del modelo y/o permitir la comparación con otros métodos.
- **Implementación del modelo.** Aplicación del modelo en un entorno real. Esto puede dar lugar a cambios en función de los resultados y también puede crear un bucle de retroalimentación al principio del proceso.

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Teniendo en cuenta las fases del proceso de ML/IA descritas anteriormente, adoptamos la clasificación de sesgos de (Suresh y Guttag, 2020). En concreto, identifican las siguientes categorías de sesgos: sesgos históricos, de representación, de medición, de agregación, de aprendizaje, de evaluación y de implementación. En las siguientes subsecciones, definimos los sesgos mencionados anteriormente y ofrecemos estudios de casos de las fuentes recopiladas para el proyecto.

Sesgo histórico

El sesgo histórico se corresponde con el sesgo preexistente tal y como lo definen Friedman y Nissenbaum (1996), que incorpora los prejuicios y estereotipos ya existentes en los datos. Se puede ver un ejemplo en Calderone (1990), que examina si la frecuencia del dolor y la medicación sedante administrada a los pacientes postoperatorios de bypass coronario (CABG) difiere según el sexo y la edad del paciente. El resultado reveló que a los pacientes varones y a los pacientes de 61 años o menos se les administraban analgésicos con mucha más frecuencia que a las pacientes mujeres y a los pacientes de 62 años o más, a quienes, en cambio, se les administraban sedantes con mucha más frecuencia. El estudio de caso sobre la precisión predictiva de los modelos de predicción del riesgo de accidente cerebrovascular en poblaciones negras y blancas se muestra en la subsección Sesgo preexistente; sin embargo, vamos a presentar otro estudio de caso que muestra el sesgo histórico en relación con el uso de la IA en la salud mental.

Estudio de caso: La inteligencia artificial en la salud mental y los sesgos de los modelos basados en el lenguaje

(Straw y Callison-Burch, 2020) presentan una revisión bibliográfica sistemática sobre los usos del PLN (procesamiento de lenguaje natural) en la salud mental, con el

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

objetivo de identificar cómo estos sesgos pueden ampliar las desigualdades en materia de salud. Los modelos de IA que utilizan el PLN para perfilar la salud mental recopilan grandes conjuntos de datos de lenguaje expresivo, normalmente obtenidos de las redes sociales, foros en línea, blogs y salas de chat. Sin embargo, estos datos ya están influenciados por los antecedentes personales y el contexto social de cada individuo.

Concretamente, en lo que respecta al género y el lenguaje, existe una extensa bibliografía (sobre el idioma inglés) resumida en (Pennebaker et al., 2003), que revela diferencias en el uso de las palabras por parte de mujeres y hombres. Por ejemplo, las mujeres utilizan un lenguaje menos asertivo, lo que se manifiesta en una mayor cortesía, menos palabrotas, más intensificadores (por ejemplo, «realmente», «tan») y más evasivas (es decir, calificativos o palabras inciertas como «más o menos», «quizás» o «tal vez»). Los hombres, por su parte, se describen como directos, precisos y también menos emocionales en su uso del lenguaje, que se caracteriza por referencias a la cantidad, adjetivos valorativos (por ejemplo, «bueno», «tonto»), frases elípticas («Gran foto») y referencias al «yo». Como señalan los autores, estas diferencias son coherentes con un marco sociológico de diferencias de género, pero también pueden atribuirse a explicaciones alternativas, como el mayor compromiso social de las mujeres.

En lo que respecta a la salud mental, los hombres y las mujeres escriben notas de suicidio en las que expresan su angustia de forma diferente; las mujeres interiorizan las emociones negativas, mientras que los hombres expresan una ira creciente (Straw y Callison-Burch, 2020). Un sistema de IA que detecta problemas de salud mental para un género puede ser inadecuado para otro (y esto es considerando el género en un contexto binario, lo que excluye a una gran parte de la población).

Sesgo de representación

El sesgo de representación se produce cuando la muestra de desarrollo subrepresenta a alguna parte de la población durante la fase de recopilación de datos. Esto puede

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

surgir de las siguientes maneras: al definir la población objetivo, si no refleja la población usuaria; al definir la población objetivo, si contiene grupos subrepresentados; al tomar muestras de la población objetivo, si el método de muestreo es limitado o desigual. El sesgo de representación da lugar a una generalización inadecuada para un subconjunto de la población usuaria. Un ejemplo típico de sesgo de representación es la detección del cáncer de piel, ya que muchos conjuntos de datos de imágenes infra-representan a determinados grupos demográficos, lo que hace que los modelos de aprendizaje automático se entrenen con imágenes de personas principalmente de piel clara (Guo et al., 2021). Teniendo en cuenta las enfermedades objetivo del proyecto AEQUITAS, presentamos un estudio de caso sobre el sesgo de representación en relación con la raza en la diabetes tipo 2.

Estudio de caso: Evaluación del sesgo racial en los algoritmos de predicción del riesgo de diabetes tipo 2

Según (Cronjé et al., 2023), en lo que respecta a la población estadounidense, a pesar de su riesgo comparativamente menor, los grupos blancos no hispanos siguen estando sobrerrepresentados en la literatura sobre predicción del riesgo de diabetes. En otra revisión sobre la equidad étnico-racial en la inteligencia artificial para el tratamiento de la diabetes, en los artículos revisados que informaban sobre la raza, la distribución media era del 69,5 % de personas blancas, el 17,1 % de negras y el 3,7 % de asiáticas, mientras que solo dos artículos informaban de la inclusión de participantes nativos americanos (Pham et al., 2021).

Está bien documentado que las desigualdades en los resultados de la diabetes se deben en gran medida a determinantes sociales de la salud complejos e interrelacionados, como el acceso a alimentos saludables, la calidad de la atención sanitaria, la situación en materia de seguros, las barreras educativas y las diferencias en las tasas de adopción tecnológica. Estos resultados incluyen tasas más altas de

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

complicaciones y un peor control glucémico entre las poblaciones minoritarias y de bajos ingresos (Alipour y Alipour, 2025).

Como resultado, un sistema de IA entrenado con conjuntos de datos existentes generalizaría de forma incorrecta, lo que daría lugar a modelos predictivos sesgados que podrían favorecer a personas de determinados grupos raciales, por ejemplo, en las medidas preventivas.

Sesgo de medición

El sesgo de medición se produce al elegir, recopilar o calcular las características y etiquetas que se utilizarán en un problema de predicción, especialmente cuando se utiliza un proxy (una aproximación de un constructo que no está codificado ni es observable directamente). Un ejemplo de ello lo encontramos en un estudio de Obermeyer et al. (2019), en el que se utilizaron los costes sanitarios como proxy para predecir y clasificar qué pacientes se beneficiarían más de una atención adicional, lo que dio lugar a una discriminación racial. Sin embargo, los costes sanitarios son un proxy deficiente de las necesidades sanitarias, ya que los y las pacientes negras, que se enfrentan a niveles desproporcionados de pobreza, suelen gastar menos en atención sanitaria que las blancas. Debido a este sesgo, el algoritmo concluyó erróneamente que las pacientes negras eran más sanas que las pacientes blancas con la misma enfermedad, por lo que las clasificó como pacientes de menor prioridad a la hora de acceder a los servicios de atención sanitaria.

Pueden darse otras fuentes de sesgo de medición cuando el método de medición varía entre los grupos, por ejemplo, cuando se supervisa el mismo comportamiento en dos grupos, pero uno de ellos se supervisa de forma más estricta o frecuente que el otro. Del mismo modo, la precisión de la medición puede variar entre los grupos, lo que en aplicaciones médicas puede dar lugar a tasas sistemáticamente más altas de diagnósticos erróneos o infradiagnósticos en determinados grupos. Por ejemplo, los médicos tienden a subestimar el dolor de los y las pacientes negras en comparación

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

con aquellas personas no negras debido a creencias erróneas sobre las diferencias biológicas entre las personas negras y las blancas, lo que hace que las primeras sean menos propensas a recibir analgésicos y, si los reciben, en cantidades menores (Hoffman et al., 2016).

Estudio de caso: Diferencias raciales y étnicas en la asociación entre la glucosa promedio y la hemoglobina A1c

La prueba de A1C mide la cantidad promedio de glucosa (azúcar) en la sangre y se utiliza para detectar la prediabetes o ayudar a diagnosticar la diabetes tipo 2. Sin embargo, la A1C es solo una medida indirecta y no está relacionada causalmente con los resultados de salud, ya que hay numerosas formas en que la relación entre las medidas directas de la glucemia (la concentración de glucosa en la sangre) y la A1C puede alterarse directamente. Existe incluso una variación sustancial en la relación entre la glucemia y la A1C entre individuos e incluso dentro de un mismo individuo a lo largo del tiempo. Además, algunos estudios han informado de niveles significativamente más altos de hemoglobina A1c (A1C) en pacientes afroamericanos que en pacientes blancos con la misma glucosa media (Karter et al., 2023).

Si un sistema de IA diseñado para diagnosticar la diabetes se entrena para utilizar los resultados de las pruebas de A1C como indicador de la glucemia sin tener en cuenta otros factores, como la raza del paciente, esto puede dar lugar a diagnósticos prematuros de diabetes y a tratamientos inadecuados, lo que se traduce en una calidad asistencial sesgada y en desigualdades en materia de salud. Sin embargo, como se señala en (Alipour y Alipour, 2025), una revisión sistemática de los sesgos que podrían afectar a la equidad de los modelos de IA/ML en la diabetes (incluido el sesgo de medición), aunque los estudios revisados mencionan explícitamente que el sesgo de medición puede propagarse a través de los modelos de IA si no se corrige, ninguno de ellos tuvo en cuenta dichos sesgos durante el desarrollo del modelo, los mitigó

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

explícitamente o informó de la corrección de las diferencias en la precisión de la medición.

Sesgo de agregación

El sesgo de agregación se produce cuando se utiliza un modelo único para un conjunto de datos que incluye grupos diversos de personas o cosas.

Podemos considerar el ejemplo de asignar datos de entrada (por ejemplo, los ingresos de una persona) a etiquetas que los describen (por ejemplo, bajos, medios, altos) y que se supone que son coherentes en todos los subconjuntos de los datos. En realidad, los antecedentes o la cultura de una persona pueden cambiar el significado real de esos números. Por ejemplo, unos ingresos «altos» en una pequeña localidad rural o en un país de ingresos bajos o medios pueden significar algo muy diferente a lo que significan en una gran ciudad o en un país de ingresos altos.

Estudio de caso: Herramientas digitales de salud para el seguimiento pasivo de la depresión

El uso de herramientas digitales para medir variables fisiológicas y conductuales para el seguimiento pasivo de la depresión se aborda en (De Angel et al., 2022), una revisión sistemática sobre el tema. Los artículos revisados examinaron las asociaciones entre la depresión y los datos conductuales objetivos obtenidos de los sensores de teléfonos inteligentes y dispositivos portátiles. Estos datos se plasmaron en características utilizadas por los modelos de IA para realizar predicciones, correspondientes al sueño, la actividad física, el ritmo circadiano, la sociabilidad, la ubicación y el uso del teléfono.

Sin embargo, los autores subrayan la heterogeneidad que surge de la diversidad de métodos utilizados para crear estas características. Por ejemplo, la característica «calidad del sueño» puede definirse midiendo el número de despertares, el número total de minutos despierto o la proporción de tiempo despierto frente al tiempo dormido en una sesión de sueño, mientras que también debemos tener en cuenta las

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

diferencias en la forma en que los sensores de los distintos dispositivos describen un evento como «sueño». Dado que todas las diferenciaciones anteriores no se tienen en cuenta y se agrupan colectivamente como «calidad del sueño», y dado que un conjunto de datos puede proceder de personas o grupos con diferentes antecedentes, culturas o normas, esta característica puede tener un significado diferente para cada uno de estos grupos o individuos.

La agregación de estos datos en una única característica puede dar lugar a un sistema que no se adapte a ningún grupo o que privilegie a la población dominante si también existe un sesgo de representación. Por ejemplo, hay pruebas de que existen diferencias de género en el sueño entre hombres y mujeres, mientras que estas últimas suelen estar infrarrepresentadas en las investigaciones sobre el sueño. Además, otros factores que no se suelen tener en cuenta en los patrones y trastornos del sueño son no distinguir el género como constructo social del sexo biológico y no considerar las identidades interseccionales definidas por la edad, la raza y la clase socioeconómica (Lok et al., 2024).

Sesgo de aprendizaje

El sesgo de aprendizaje surge cuando las opciones de modelización amplifican las disparidades de rendimiento entre los diferentes ejemplos de los datos. Un ejemplo es la privacidad diferencial, un mecanismo utilizado en los sistemas de IA que garantiza que, al examinar los resultados de un sistema, no sea posible determinar si los datos de una persona concreta se incluyeron en el conjunto de datos original. La privacidad diferencial se utiliza en los conjuntos de datos sanitarios para proteger la información confidencial de los pacientes, por ejemplo, en el caso de enfermedades raras, en las que el caso de cada paciente es más o menos único en un área limitada cubierta por un hospital, por lo que, incluso si los datos se anonimizan, no es muy difícil deducir la identidad de la persona. Sin embargo, se ha demostrado que la privacidad diferencial reduce la influencia de los datos infrarrepresentados en el modelo; por lo tanto, si el

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

sistema de IA está sesgado desde el principio, la aplicación de una medida de mejora de la privacidad exacerba aún más este sesgo (Bagdasaryan y Shmatikov, 2019).

Estudio de caso: Privacidad diferencial y disparidades en materia de salud

En septiembre de 2018, la Oficina del Censo de los Estados Unidos anunció que implementaría la privacidad diferencial en los productos de datos derivados de los datos del censo de 2020. Sin embargo, (Santos-Lozada et al., 2020) investigaron la forma en que la implementación de la privacidad diferencial puede alterar el conocimiento sobre las disparidades en materia de salud en la mortalidad, especialmente para las minorías raciales o étnicas en áreas pequeñas y entornos menos urbanos. Sus resultados sugirieron que la privacidad diferencial afectará más fuertemente a las estimaciones de la tasa de mortalidad de las personas afroamericas no hispanas y de las hispanas, que a las estimaciones de las personas blancas no hispanas.

Estos hallazgos fueron respaldados por (Kurz et al., 2022), quienes muestran que la aplicación de la privacidad diferencial a los mismos datos puede dar lugar a una representación errónea de las tasas de participación en Medicaid entre los grupos raciales y étnicos ya marginados. Concretamente, estas tasas para determinadas combinaciones de condado, raza y etnia diferían entre los resultados de los datos de privacidad diferencial y los datos originales, superando en ocasiones el 10 %. Además, las personas blancas no hispanas eran el único subgrupo étnico y racial para el que el algoritmo de privacidad diferencial captaba con precisión las tasas de participación en Medicaid. Este hallazgo puede tener importantes implicaciones para la política sanitaria, ya que los datos del censo se utilizan para planificar programas gubernamentales, asignar recursos y evaluar y realizar un seguimiento de las políticas.

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanciado por la Unión Europea

AEQUITAS

Sesgo de evaluación

El sesgo de evaluación se produce cuando los datos de referencia utilizados para una tarea concreta no representan a la población usuaria. Los puntos de referencia son conjuntos de datos estandarizados que se utilizan para medir la calidad de un modelo, lo que permite realizar comparaciones cuantitativas entre modelos. En consecuencia, existe el riesgo de fomentar el desarrollo y la implementación de modelos que solo funcionan bien en el subconjunto de datos representados en el punto de referencia. Por lo tanto, si el punto de referencia está sujeto a sesgos históricos, representativos o de medición, puede producirse una discriminación contra subgrupos o individuos vulnerables.

En el ámbito de la atención sanitaria, las razones de la infrarrepresentación de poblaciones específicas en los conjuntos de datos pueden deberse a que las personas o los grupos están ausentes de los conjuntos de datos (por ejemplo, las mujeres embarazadas, debido a restricciones éticas) o a que las personas están clasificadas de forma incorrecta o inadecuada en grupos (por ejemplo, las categorías de «etnia mixta» u «otros»). Las causas fundamentales de esto pueden incluir razones sociales y técnicas o legales/éticas, como barreras estructurales para recibir atención sanitaria, obstáculos técnicos para la captura o digitalización de datos sanitarios relevantes, limitaciones individuales y estructurales en relación con el consentimiento para compartir datos, y restricciones legales o éticas sobre el intercambio de datos que impiden la accesibilidad de los mismos, entre otras (Arora et al., 2023). El resultado es que los sistemas de IA calibrados según esos puntos de referencia pueden tener un rendimiento inferior cuando se aplican a personas de un grupo infrarrepresentado. Sin embargo, es importante señalar que la validez de los puntos de referencia es una cuestión más genérica y no se limita al sesgo (Brooks, 2025).

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanciado por la Unión Europea

AEQUITAS

Estudio de caso: Conjuntos de datos de imágenes de la piel

Los conjuntos de datos de imágenes de la piel infrarrepresentan a ciertos grupos demográficos, ya que la mayoría de las imágenes de estos conjuntos proceden de poblaciones de América del Norte o Europa y representan predominantemente a personas de piel clara (Guo et al., 2021). Debido al alto coste y la dificultad de construir estos conjuntos de datos, además de para entrenar modelos, también pueden utilizarse como puntos de referencia.

El caso práctico que ilustra el sesgo emergente, es decir, los conjuntos de datos de imágenes de cáncer de piel utilizados para entrenar modelos de predicción, es un ejemplo de referencia inadecuada cuando la población de usuarios proviene de grupos infrarrepresentados (Guo et al., 2021). Un caso similar, aunque no relacionado con la IA, muestra la generalidad del problema, que tenía que ver con los oxímetros de pulso (dispositivos que miden la saturación de oxígeno en sangre, utilizados, por ejemplo, en casos de infarto o insuficiencia cardíaca), que han demostrado funcionar con mayor precisión en pieles de pigmentación clara (Sjoding et al., 2020).

Los sesgos de representación, medición, agregación, aprendizaje y evaluación pueden asignarse al sesgo técnico definido por (Friedman y Nissenbaum, 1996).

Sesgo de implementación

El sesgo de implementación surge cuando existe un desajuste entre el problema que un modelo pretende resolver y la forma en que se utiliza realmente, lo que puede causar daños, especialmente cuando se combina con sesgos cognitivos como los sesgos de confirmación y automatización. El sesgo de implementación es lo mismo que el sesgo emergente definido por (Friedman y Nissenbaum, 1996).

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Caso práctico: Cambio de dominio

El caso del cambio de datos se documenta en la subsección sesgo emergente sobre la detección del cáncer de piel. Además, podemos definir el caso del cambio de dominio, que se produce cuando un sistema se implementa, ha superado la autorización reglamentaria y se utiliza en la práctica clínica, pero se aplica a una cohorte de pacientes diferente a aquella para la que se entrenó. Por ejemplo, se puede desarrollar un sistema para un hospital de un país de ingresos altos y desplegarlo en un país de ingresos bajos o medios sin tener en cuenta factores como las características sociodemográficas de los pacientes o si estos tienen el mismo nivel de riesgo general que los incluidos en los datos de entrenamiento (Vokinger et al., 2021).

Implicaciones políticas

Las evidencias recogidas en el Deliverable D2.1 demuestran que los sesgos de género y raciales en la IA biomédica no son defectos técnicos incidentales o aislados, sino riesgos sistémicos que surgen a lo largo de todo el ciclo de vida de los sistemas de IA utilizados en la asistencia sanitaria. En las enfermedades cardiovasculares, la depresión y la diabetes, los sesgos se derivan de conjuntos de datos clínicos históricamente sesgados, prácticas diagnósticas desiguales, variables proxy que codifican desigualdades estructurales y contextos de implementación que distribuyen de forma desigual tanto los beneficios como los perjuicios. Estos hallazgos confirman que la IA biomédica afecta directamente a múltiples derechos y principios protegidos por la Carta de los Derechos Fundamentales de la Unión Europea, en particular los preceptos de dignidad humana, igualdad ante la ley y no discriminación, así como el derecho a la integridad de la persona, el derecho a la asistencia sanitaria, la protección de datos y el derecho a un recurso efectivo.

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanciado por la Unión Europea

AEQUITAS

En este contexto, los marcos políticos de la UE y nacionales que regulan la IA en la asistencia sanitaria deben tratar la mitigación de los sesgos no como un complemento ético voluntario, sino como un componente vinculante del despliegue de la IA conforme a la ley y a los derechos. Los esfuerzos reguladores europeos y nacionales en materia de IA en la asistencia sanitaria deben considerarse dentro del marco más amplio de los derechos fundamentales que rigen la IA (véase Novossiolova, 2025; Novossiolova et al., 2025; Kasapi, 2025). El Reglamento Europeo de IA proporciona una base normativa necesaria al clasificar la mayor parte de la IA biomédica como sistemas de alto riesgo, pero su eficacia en la práctica dependerá de cómo se apliquen las garantías de los derechos fundamentales en las evaluaciones de conformidad, la supervisión posterior a la comercialización y la contratación pública.

En primer lugar, deben reforzarse y especificarse las garantías de una supervisión humana significativa para los sistemas de IA biomédica a lo largo de todo su ciclo de vida. Las herramientas clínicas de IA utilizadas para el diagnóstico, la estratificación del riesgo, el cribado o el apoyo al tratamiento no deben, en ningún caso, funcionar como responsables de la toma de decisiones autónomos de facto. La supervisión humana debe incluir no solo la posibilidad de que los profesionales sanitarios anulen las decisiones, sino también una responsabilidad institucional clara para comprender las limitaciones del sistema, los riesgos de sesgo conocidos y las diferencias de rendimiento entre subgrupos. En consonancia con la protección de la dignidad y la integridad humanas que recoge la Carta, los profesionales sanitarios deben recibir formación y apoyo institucional para cuestionar críticamente los resultados de la IA en lugar de limitarse a aceptarlos. Esto requiere incorporar la alfabetización en IA, la concienciación sobre los sesgos y la formación en derechos fundamentales en la formación médica y el desarrollo profesional continuo.

Las obligaciones de transparencia deben interpretarse de manera amplia en los contextos sanitarios. Los pacientes y los usuarios de los servicios sanitarios deben ser informados siempre que se utilicen sistemas de IA en la toma de decisiones clínicas

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

que les afecten, incluyendo el cribado, la priorización o la evaluación de riesgos. Cuando los resultados generados por la IA sirvan de base para los servicios de salud pública, dichos resultados deben ser claramente identificables como tales y acompañarse de explicaciones accesibles sobre su función, sus limitaciones y los riesgos de sesgo conocidos. También se debe informar a las personas cuando sus datos personales se utilicen para el entrenamiento, las pruebas o el aprendizaje continuo de la IA, especialmente cuando se trate de datos sanitarios sensibles. Estas medidas de transparencia son esenciales para defender los derechos de la Carta en materia de protección de datos y recursos efectivos, y para que las personas puedan impugnar de manera significativa las decisiones que puedan afectarles negativamente.

En segundo lugar, la evaluación del impacto sobre los derechos fundamentales debe convertirse en un requisito rutinario y exigible para los sistemas de IA biomédicos, que vaya más allá de los controles previos a la comercialización y se extienda a la evaluación continua durante su implementación. Las pruebas empíricas del documento D2.1 muestran que muchos daños por sesgos solo se hacen visibles una vez que los sistemas de IA interactúan con poblaciones reales y flujos de trabajo clínicos, en particular a través de efectos interseccionales que implican el género, la raza, la edad y la situación socioeconómica. Por lo tanto, las evaluaciones de impacto basadas en los derechos, como las inspiradas en la metodología HUDERIA del Consejo de Europa (Metodología para la evaluación de riesgos e impacto de los sistemas de inteligencia artificial desde el punto de vista de los derechos humanos, la democracia y el Estado de derecho), deberían ser obligatorias para la IA médica de alto riesgo, examinando explícitamente el rendimiento y los resultados diferenciales entre los grupos protegidos. Estas evaluaciones deben contar con la participación significativa de las partes interesadas, incluidas las organizaciones de la sociedad civil, los representantes de los pacientes y los organismos de igualdad, con el fin de poner de manifiesto los daños que pueden ser invisibles desde una perspectiva puramente técnica o clínica.

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Deben exigirse auditorías periódicas de los sistemas de IA biomédica para verificar el cumplimiento continuo de las normas de derechos fundamentales, prestando especial atención a la deriva de los sesgos, los cambios en los conjuntos de datos y los cambios en el uso clínico a lo largo del tiempo. Cuando las auditorías revelen efectos discriminatorios persistentes o imposibles de mitigar, debe haber vías legales e institucionales claras para restringir, suspender o poner fin al uso del sistema. El derecho a la asistencia sanitaria no puede justificar el despliegue continuado de herramientas de IA que perjudican sistemáticamente a determinados grupos, incluso si las métricas de rendimiento agregadas parecen favorables.

En tercer lugar, las autoridades de la UE y nacionales deben abordar el riesgo de uso indebido y los daños secundarios asociados a la IA biomédica. Esto incluye las vulnerabilidades de ciberseguridad que podrían comprometer la integridad del sistema o permitir la manipulación maliciosa de los resultados clínicos, así como la reutilización de la IA sanitaria para prácticas de vigilancia, elaboración de perfiles o exclusión. Los sistemas de IA biomédica deben estar sujetos a evaluaciones de seguridad periódicas y a obligaciones estrictas de notificación de incidentes, con mecanismos claros de rendición de cuentas en los casos en que los sistemas sesgados o comprometidos den lugar a violaciones de derechos. Los marcos de responsabilidad deben garantizar que la responsabilidad no se pueda desviar únicamente hacia los médicos individuales cuando los daños están estructuralmente integrados en el diseño de la IA o en las decisiones de implementación.

En cuarto lugar, la promoción de prácticas éticas y responsables debe integrarse en toda la cadena de valor de la IA biomédica. Se debe exigir a los desarrolladores que aborden de forma proactiva los riesgos de sesgo mediante la recopilación de datos representativos, la selección cuidadosa de objetivos y proxies, la validación específica de subgrupos y la presentación de informes transparentes sobre el rendimiento en todos los grupos de género y raza. Es importante destacar que las evidencias revisadas en D2.1 muestran que la «equidad por desconocimiento» y las estrategias de

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

eliminación de sesgos puramente técnicas suelen ser insuficientes en el ámbito sanitario. Por lo tanto, las directrices y normas reglamentarias deben ir más allá de las métricas abstractas de equidad y exigir a los desarrolladores que demuestren resultados de equidad clínicamente significativos, evaluados en relación con las vías de atención sanitaria reales y los patrones de acceso.

Las políticas de contratación pública y financiación desempeñan un papel crucial en la configuración de los incentivos para los desarrolladores. Las autoridades sanitarias y los hospitales públicos deben integrar los derechos fundamentales y los criterios de sesgo en las decisiones de contratación de sistemas de IA, favoreciendo las soluciones que demuestren prácticas de mitigación del sesgo sólidas, transparentes y verificadas de forma independiente. Los instrumentos de financiación de la UE, incluidos los futuros programas de investigación e innovación, deben seguir dando prioridad a los proyectos que combinan la innovación técnica con la gobernanza basada en los derechos, la participación de las partes interesadas y el desarrollo de capacidades, en línea con el modelo AEQUITAS.

Por último, el fortalecimiento de la resiliencia social frente a la IA biomédica sesgada requiere una inversión sostenida en la sensibilización pública, la participación de la sociedad civil y la colaboración intersectorial. Se debe empoderar a las personas para que comprendan sus derechos en la asistencia sanitaria mediada por la IA y los mecanismos disponibles para protegerlos. Las organizaciones de la sociedad civil, los organismos de igualdad y los grupos de pacientes deben ser reconocidos como actores esenciales en la supervisión de los impactos de la IA, el apoyo a las personas afectadas y la información sobre el desarrollo de políticas. La cooperación entre los gobiernos, los proveedores de asistencia sanitaria, los investigadores, la industria y la sociedad civil es necesaria para garantizar que los beneficios de la IA biomédica se compartan de forma equitativa y no refuercen las desigualdades sanitarias existentes.

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

En conjunto, las conclusiones del Deliverable D2.1 respaldan una conclusión política clara: la IA biomédica solo puede considerarse fiable y legítima en la UE cuando su diseño, despliegue y gobernanza están firmemente anclados en la protección de los derechos fundamentales. El Reglamento Europeo de IA, interpretado a través del prisma de la Carta de los Derechos Fundamentales de la UE y puesto en práctica mediante mecanismos concretos de supervisión, evaluación de impacto y rendición de cuentas, ofrece una oportunidad crucial para garantizar que la innovación en la asistencia sanitaria promueva la equidad en lugar de reproducir los patrones históricos de discriminación.

Financiado por la Unión Europea. Las opiniones y puntos de vista expresados solo comprometen a su(s) autor(es) y no reflejan necesariamente los de la Unión Europea o los de la Agencia Ejecutiva Europea de Educación y Cultura (EACEA). Ni la Unión Europea ni la EACEA pueden ser considerados responsables de ellos. Código de proyecto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI