Universidad Internacional de La Rioja
Escuela Superior de Ingeniería y Tecnología

Grado en Ingeniería Informática

Sistema de Gestión de Bases de Datos para el Estudio de la Utilización de Medicamentos en un Hospital

Trabajo fin de estudio presentado por:	Vicente Ribera Damiá
Director/a:	Claudia Blanca González Calleros
Fecha:	16 de julio de 2025
Repositorio del código fuente:	TFE-Polimedicados

Resumen

La polifarmacia, entendida como el uso simultáneo de cinco o más medicamentos, plantea importantes desafíos para la seguridad del paciente, especialmente en personas de edad avanzada. Aumenta el riesgo de reacciones adversas, interacciones farmacológicas y hospitalizaciones, lo que repercute negativamente en la salud del paciente y en los costos del sistema sanitario. En este contexto, la vigilancia activa de las reacciones adversas a medicamentos es clave para garantizar una atención segura y eficaz.

Este trabajo se centró en el desarrollo de una solución tecnológica para el estudio de reacciones adversas a medicamentos en pacientes polimedicados de edad avanzada, mediante la creación de un repositorio sintético anonimizado, basado en datos simulados que reflejan las características reales del servicio de urgencias de un hospital de tamaño medio.

La metodología aplicada abarcó desde el análisis de las necesidades de la unidad de farmacología del hospital hasta la implementación local del repositorio, siguiendo un enfoque de desarrollo iterativo que permitió ajustes y mejoras continuas, garantizando una solución adaptable, escalable y alineada con los objetivos del proyecto.

Los resultados obtenidos demuestran que el uso de datos sintéticos anonimizados ha sido muy útil para que los investigadores puedan analizar RAM en pacientes polimedicados mayores sin comprometer datos reales. La solución permitió validar el diseño técnico, ejecutar consultas analíticas y detectar combinaciones de riesgo, aunque con ciertas limitaciones en la fidelidad de los datos. Además, ha acelerado el desarrollo y facilitado la futura implementación con datos reales.

En resumen, esta solución permite avanzar en el estudio de ingresos hospitalarios por RAM en pacientes polimedicados, al identificar combinaciones de fármacos potencialmente evitables. Su integración con tecnologías como la inteligencia artificial abre nuevas oportunidades para mejorar la seguridad del paciente y apoyar la investigación.

Palabras clave: Farmacovigilancia, Polimedicados, Reacciones adversas a medicamentos, Coste, Ingresos hospitalarios.

Abstract

Polypharmacy, understood as the simultaneous use of five or more medications, poses significant challenges for patient safety, especially in older adults. It increases the risk of adverse reactions, drug interactions, and hospitalizations, negatively impacting patient health and healthcare system costs. In this context, active monitoring of adverse drug reactions is key to ensuring safe and effective care.

This work focused on developing a technological solution for studying adverse drug reactions in elderly patients taking multiple medications. This approach focused on creating an anonymized synthetic repository based on simulated data that reflects the real-life characteristics of a medium-sized hospital's emergency department.

The methodology applied ranged from analyzing the needs of the hospital's pharmacology unit to implementing the repository locally. The solution followed an iterative development approach that allowed for continuous adjustments and improvements, ensuring an adaptable, scalable solution aligned with the project's objectives.

The results obtained demonstrate that the use of anonymized synthetic data has been very useful in enabling researchers to analyze ADRs in older polymedicated patients without compromising real-world data. The solution allowed for the validation of the technical design, the execution of analytical queries, and the detection of risky combinations, albeit with certain limitations in data accuracy. Furthermore, it has accelerated development and facilitated future implementation with real-world data.

In summary, this solution allows for progress in the study of hospital admissions due to ADRs in polymedicated patients by identifying potentially avoidable drug combinations. Its integration with technologies such as artificial intelligence opens up new opportunities to improve patient safety and support research.

Keywords: Pharmacovigilance, Polymedications, Adverse drug reactions, Cost, Hospital admissions.

Índice de contenidos

Índice de figuras

Índice de tablas

1. Introducción

El envejecimiento poblacional y el aumento de la esperanza de vida han producido un incremento significativo en la carga de enfermedades crónicas y en la polimedicación (Maher et al., 2014). Este fenómeno, además de su impacto en la calidad de vida del paciente, supone un desafío clínico y sanitario. La polimedicación, definida comúnmente como el uso de cinco o más medicamentos simultáneos, se asocia con un mayor riesgo de reacciones adversas a los medicamentos (RAM), errores de medicación y un incremento de ingresos hospitalarios provocados por RAM. Ante esta situación, la vigilancia activa de RAM se convierte en una herramienta crucial para promover la seguridad del paciente y la eficacia terapéutica.

Desde una perspectiva asistencial, la identificación precoz de RAM permite prevenir complicaciones, reducir hospitalizaciones y ajustar tratamientos farmacológicos. Además, el análisis de patrones de RAM en poblaciones de pacientes geriátricos y polimedicados, puede ayudar a disminuir el riesgo de interacciones medicamentosas que puedan derivar en complicaciones clínicas graves (Zazzara et al., 2021).

En conjunto, las motivaciones clínicas, demográficas y operativas justifican el desarrollo de herramientas destinadas a la predicción de RAM en pacientes polimedicados de edad avanzada. Este trabajo explora cómo la implementación de un repositorio externo anonimizado enfocado en la polimedicación puede contribuir en la detección proactiva de alertas tempranas cuando se cruzan datos de laboratorio, diagnósticos y prescripciones para descubrir RAM evitables y riesgos específicos en personas de edad avanzada.

Esta solución, además de abordar desafíos técnicos relacionados con la manipulación de grandes cantidades de datos, pone de relieve las complejidades éticas, legales y prácticas de la gestión de información real, por lo que se optó por emplear datos anonimizados para proteger la privacidad de los pacientes y poder llevar a cabo estudios retrospectivos sobre la base de un repositorio sintético con datos anonimizados. Además, los avances en tecnologías emergentes como la inteligencia artificial (IA) y otras más consolidadas como el *Big Data* adaptadas al entorno hospitalario, pueden contribuir en la mejora de una práctica médica más segura y, por consiguiente, reducir la prevalencia de ingresos hospitalarios provocados por RAM en pacientes polimedicados de edad avanzada.

1.1. Motivación

La elección de este tipo de trabajo surge por un interés personal sobre los problemas que afectan directamente a la salud y a la vida de las personas mayores que nos rodean, sobre todo cuando tus seres queridos van entrando en años, incrementándose las visitas al hospital y la cantidad de medicamentos prescritos para curar o aliviar sus enfermedades, siendo algunas de ellas crónicas y para las que ya se están tratando, lo que se traduce en un solapamiento de fármacos que puede derivar en un ingreso hospitalario provocado por una RAM.

El haber vivido en un entorno médico desde la niñez y habiéndome formado en el área de la informática, me ha motivado a afrontar este desafío tan apasionante empleando mis habilidades y conocimientos técnicos para contribuir significativamente en un ámbito tan crucial como la medicina preventiva centrada en reducir el riesgo de sufrir una RAM, la utilización adecuada de los medicamentos y la farmacovigilancia.

Los sistemas de farmacovigilancia hospitalaria velan por la seguridad del paciente y por el uso racional de los medicamentos para mejorar la calidad de vida de las personas. Sin embargo, presentan diversas limitaciones: tienden a centrarse en la priorización de señales individuales de RAM, dejando de lado aquellas derivadas de combinaciones complejas de fármacos, lo que dificulta la identificación de RAM graves o inesperadas. Además, suelen carecer de un enfoque geriátrico que contemple adecuadamente la polifarmacia en pacientes mayores. A esto se suma la fragmentación de los datos clínicos, dispersos en distintos sistemas dentro del hospital, lo que puede retrasar la identificación de interacciones medicamentosas y entorpecer el análisis oportuno de señales de seguridad.

En este contexto, resulta imprescindible contar con conocimientos técnicos que permitan incorporar herramientas complementarias de análisis de datos clínicos. Estas herramientas pueden facilitar la detección de patrones de riesgo que ayuden a optimizar los tratamientos y prevenir ingresos hospitalarios asociados a RAM, especialmente en pacientes mayores polimedicados. Todo ello supone un desafío considerable para hospitales de tamaño medio, donde los recursos suelen ser limitados.

1.2. Planteamiento del trabajo

Este trabajo presenta una solución tecnológica diseñada específicamente para dar respuesta a las necesidades de los profesionales de la Unidad de Farmacología (UF) de un hospital de tamaño medio. Su objetivo es facilitar la realización de estudios sobre RAM en pacientes polimedicados de edad avanzada, un grupo especialmente vulnerable. Además, la herramienta garantiza la protección de la privacidad de los pacientes mediante el uso de datos previamente anonimizados, cumpliendo con los estándares éticos y legales vigentes.

La aportación más relevante de este trabajo ha consistido en el diseño e implementación de un repositorio sintético con datos anonimizados, desarrollado en paralelo, pero de forma coordinada con el Departamento de Sistemas de Información (DSI) del hospital. Este repositorio utiliza datos anonimizados desde su origen, lo que permite desacoplar su desarrollo de los sistemas del hospital, simular la salida de un proceso de Extracción, Transformación y Carga (ETL), generando datos sintéticos con estructuras idénticas a las de los datos reales. Esto posibilita validar reglas de anonimización, probar flujos de trabajo sin depender del acceso a datos reales, reducir los riesgos asociados al manejo de información confidencial y de ese modo poder validar la utilidad de los datos. En conjunto, esta estrategia permite acelerar la puesta en producción del repositorio definitivo con datos reales anonimizados.

Este trabajo ha tenido por objetivo desarrollar una arquitectura de procesamiento de datos clínicos sintéticos que permita su anonimización, almacenamiento y análisis en un entorno relacional, con el fin de detectar patrones asociados a RAM. Para ello, se simula un proceso ETL que transforma y anonimiza los datos mediante técnicas como k-anonimidad, generalización y supresión, preservando su utilidad analítica. Los datos transformados se cargan en un repositorio sintético anonimizado desarrollado en PostgreSQL, sobre el que se podrán ejecutar consultas orientadas a la detección de RAM en pacientes polimedicados de edad avanzada, así como a generar evidencia local sobre combinaciones farmacológicas vinculadas a hospitalizaciones evitables. Esta arquitectura facilita un entorno seguro y flexible para la realización de estudios orientados al análisis de riesgos clínicos sin comprometer información sensible.

1.3. Estructura del trabajo

El trabajo consta de varios capítulos en los que se desarrolla una base teórica y práctica, finalizando con la materialización y evaluación de una solución para la toma de decisiones en base a los estudios realizados por la UF del hospital.

En el Capítulo 2 se establece el contexto y el estado del arte, analizando la relevancia de las Reacciones Adversas de los Medicamentos (RAM). Este capítulo enmarca el trabajo dentro de la gestión de las RAM, discutiendo las necesidades y desafíos asociados con los sistemas de notificaciones de RAM y las soluciones existentes para el análisis de datos sobre RAM.

En el Capítulo 3, se exponen los objetivos, la metodología adoptada para desarrollar el repositorio sintético anonimizado empleando datos simulados, desde su conceptualización hasta su implementación final con el propósito de proporcionar una herramienta eficaz para la UF.

El Capítulo 4 se centra en la implementación de la solución, los resultados obtenidos, los desafíos enfrentados y cómo fueron solucionados. Además, se evalúa el funcionamiento de la herramienta y su efectividad como sistema para la toma de decisiones y apoyo en la investigación de RAM.

Finalmente, en el Capítulo 5 se expone una reflexión sobre los hallazgos e implicación de la solución desarrollada, concluyendo con la evaluación del impacto del proyecto, poniendo de relieve las limitaciones de la solución y proponiendo líneas de trabajo futuras de acuerdo con tecnologías emergentes complementarias como la inteligencia artificial y el aprendizaje automático para mejorar la atención al paciente, la calidad de los diagnósticos y prescripciones, la generación de alertas en tiempo real en pacientes polimedicados de edad avanzada, la aceleración de los procesos de investigación, y en general, la gestión de la salud.

2. Contexto y Estado del Arte

En esta sección se proporciona una visión general sobre la problemática y relevancia del estudio relacionado con las reacciones adversas a los medicamentos (RAMs), la seguridad en el uso de los medicamentos y el papel que desempeña la UF de un hospital de tamaño medio. Por otra parte, se pone de relieve la importancia y complejidad del problema a la hora de llevar a cabo estudios sobre la prevalencia de ingresos hospitalarios provocados por RAM, así como los principales desafíos para el acceso y procesamiento de datos clínicos, teniendo en cuenta que dichos datos deben ser fiables, precisos y de calidad, todo ello sin perder de vista las técnicas de seudonimización o anonimización de los datos como mecanismos de protección de datos clínicos cuando vayan a ser empleados para la realización de estudios sobre RAM. Además, se analizan aspectos importantes relacionados con la interpretación de los resultados y su repercusión en la seguridad del paciente.

2.1. Análisis del contexto

Este trabajo se inserta dentro del marco de la **farmacovigilancia** hospitalaria y el apoyo en sistemas de información clínica para mejorar la detección y prevención de interacciones peligrosas de los fármacos, todo ello alineado con los esfuerzos globales en materia de seguridad del paciente y uso racional de los medicamentos.

En el ámbito de la farmacovigilancia, la Organización Mundial de la Salud (OMS) define las reacciones adversas a medicamentos como “aquellas respuestas negativas e involuntarias que pueden surgir tras la administración de un fármaco a dosis habituales, utilizadas con fines terapéuticos, preventivos o diagnósticos”. Este concepto excluye efectos derivados de sobredosis o errores de medicación, centrándose en eventos indeseables que ocurren bajo un uso clínicamente apropiado del medicamento (WHO Meeting on International Drug Monitoring: the Role of National Centres (1971: Geneva, 1972).

Algunas características del paciente pueden predisponer a sufrir una RAM, como la edad, la comorbilidad (enfermedades preexistentes), interacciones farmacológicas o una predisposición genética a determinadas RAM (Sociedad Española de Farmacia Hospitalaria, 2024).

Existen dos tipos de RAM, las de tipo A que "son resultado de un aumento en la acción farmacológica del medicamento cuando se administra a la dosis terapéutica habitual" (*AEMPS*, 2015), por ejemplo, un sangrado excesivo provocado por una dosis demasiado alta de anticoagulantes o una hipoglucemia por una administración excesiva de insulina. En definitiva, podríamos decir que este tipo de RAM es predecible. Pero existe otro tipo de RAM que es impredecible y cuya sintomatología resulta más complicada de diagnosticar, como son las RAM de tipo B que son más difíciles de predecir y no son dosis-dependiente, siendo aquellas que "no se esperan y son provocadas por acciones farmacológicas del fármaco" (*AEMPS*, 2015), por ejemplo, una aplasia medular producida por un fármaco que no es dosis-dependiente y provoca un fallo irreversible de la médula ósea, o bien una erupción cutánea provocada por un antibiótico. Este último tipo de RAM representa uno de los objetivos principales de la Farmacovigilancia (*AEMPS*, 2015).

Tabla 1. Características de las reacciones Tipo A y Tipo B.

Tipo A	Tipo B
Predecibles	No predecibles
Normalmente dosis-dependiente	Raramente dosis-dependiente
Alta morbilidad	Baja morbilidad
Baja mortalidad	Alta mortalidad
Responde a una reducción de la dosis	Responde a la retirada del fármaco

*Fuente: (*AEMPS*, 2015)*

Las RAMs representan un problema significativo de salud pública, ya que pueden generar complicaciones en los pacientes, prolongar la estancia hospitalaria e incrementar los costos sanitarios (Howard et al., 2007). Se estima que un porcentaje considerable de los ingresos hospitalarios está relacionado con RAM prevenibles, lo que subraya la necesidad de

estrategias para mejorar la identificación y gestión de estos eventos (Bates et al., 2003). La seguridad en el uso de los medicamentos es un aspecto crítico en la atención hospitalaria.

Las **interacciones medicamentosas (IM)** en pacientes polimedicados pueden derivar en una modificación del efecto terapéutico por la administración simultánea de dos o más fármacos o bien en la aparición de efectos adversos no deseados. Estas interacciones pueden llegar a ser especialmente graves en pacientes de edad avanzada con enfermedades crónicas debido a la disminución de la efectividad del tratamiento, así como complicaciones en la salud del paciente, por lo que estas interacciones son relevantes en la práctica clínica y una detección precoz puede evitar o disminuir reacciones adversas en los pacientes (Bermeo Cerón et al., 2024).

Un estudio realizado por diferentes departamentos del Complejo Hospitalario A Coruña, España, sobre la "Prevalencia de reacciones adversas a medicamentos asociadas a visitas al servicio de urgencias y factores de riesgo de hospitalización", realizado del 15 de noviembre al 15 de diciembre de 2021 para determinar las variables predictoras de hospitalización por reacciones adversas a medicamentos, reveló que 10.799 pacientes visitaron el servicio de urgencias, de los que 216 (2%) presentaban reacciones adversas a medicamentos. La edad media fue de 70 ± 17,5 (18-98) años y el 47,7% de los pacientes fueron hombres. Un 54,6% de los pacientes requirieron hospitalización y el 1,6% fallecieron a causa de una RAM. El número total de fármacos involucrados fue de 315, con 149 fármacos diferentes. El número de casos (n) para el grupo farmacológico correspondiente al sistema nervioso constituyó el grupo más representativo (n = 81). Medicamentos de alto riesgo como los antitrombóticos (n = 53), fueron el subgrupo de medicamentos que causó más visitas a urgencias y hospitalizaciones. El acenocumarol (n = 20), empleado para evitar y tratar la formación de coágulos sanguíneos, es decir, evitar trombosis y embolias, fue el principal fármaco implicado. Los trastornos gastrointestinales (n = 62) fueron mayoritarios. La diarrea (n = 16) fue la reacción adversa más frecuente, mientras que la hemorragia gastrointestinal (n = 13) causó el mayor número de hospitalizaciones (Brandariz-Núñez et al., 2022).

Del estudio realizado por los distintos departamentos del hospital se reveló que la prevalencia de visitas al servicio de urgencias por reacciones adversas a medicamentos continúa siendo un problema sanitario no despreciable. Medicamentos de alto riesgo como los agentes antitrombóticos fueron el principal subgrupo terapéutico implicado. El índice de comorbilidad

de Charlson, empleado para predecir el riesgo de mortalidad a un año en pacientes con enfermedades crónicas o con más enfermedades adicionales junto a una enfermedad primaria, se comportó como un factor independiente de hospitalización, mientras que la hemorragia gastrointestinal fue la reacción adversa con mayor número de ingresos hospitalarios (Brandariz-Núñez et al., 2022).

Los **problemas relacionados con los medicamentos (PRM)** pueden afectar de manera negativa a la salud del paciente por lo que saber manejar este tipo de situaciones es importante para garantizar los tratamientos y la seguridad del paciente. Podríamos mencionar como PRM principales, las indicaciones no adecuadas cuando no existe un fármaco específico para una enfermedad, la ocurrencia de reacciones adversas, dosis inadecuadas, interacciones medicamentosas por la administración solapada de fármacos o vías de administración inadecuadas entre otras (Bermeo Cerón et al., 2024).

Uno de los principales riesgos relacionados con los PRM son las transiciones de los pacientes entre distintos centros hospitalarios, lo que provoca una desconexión entre historiales médicos de los pacientes, prescripciones inadecuadas, pérdida de información, la polimedicación a partir de un consumo de más de cinco medicamentos diarios y la falta de coordinación entre los profesionales y los pacientes (Bermeo Cerón et al., 2024).

Podemos exponer como ejemplo un caso clínico relacionado con una RAM padecida por una mujer de 71 años polimedicada que acudió al servicio de urgencias tras presentar un episodio de estupor y depresión respiratoria. Entre sus antecedentes médicos destacaban hipertensión arterial, obesidad, enfermedad pulmonar obstructiva crónica (EPOC) moderada y distimia (un tipo de depresión crónica leve).

Como parte de su tratamiento crónico la paciente estaba tomando:

- Duloxetina, un antidepresivo.
- Clorazepato dipotásico, una benzodiacepina.
- Bupropión, también antidepresivo.
- Omeprazol, para el control gástrico.
- Tramadol, un analgésico opioide para el dolor.

Debido a un control ineficaz del dolor, el facultativo de Atención Primaria decidió realizar modificaciones terapéuticas. Se sustituyó el tramadol por un parche transdérmico de

buprenorfina, un opioide más potente, y se añadió pregabalina, utilizada frecuentemente para el dolor neuropático.

Con el paso de las semanas, las dosis de buprenorfina y pregabalina fueron aumentadas con la intención de lograr un mejor alivio del dolor. No obstante, la paciente empezó a mostrar signos de somnolencia excesiva, así como una disminución de la actividad de las funciones intelectuales acompañada de falta de reacción. Fue atendida por el Servicio de Emergencias tras un episodio de pérdida de conciencia, constatándose una baja saturación de oxígeno y un estado mental alterado, por lo que tuvo que ser trasladada a Urgencias hospitalarias y suministrarle ventilación mecánica no invasiva para estabilizarla.

Al investigar la causa del deterioro, se descubrió que, además del tratamiento prescrito, la paciente había retomado su tratamiento anterior con tramadol por su cuenta y había comenzado a tomar alprazolam, un ansiolítico que le había pautado su psiquiatra de rescate. Esta combinación de opioides y benzodiacepinas provocó una depresión respiratoria grave, una RAM potencialmente letal.

Como comentario final del caso, los autores subrayan la importancia de una adecuada conciliación de la medicación, especialmente en pacientes mayores y polimedicados. Además, se remarca la necesidad de una mejor comunicación entre los distintos niveles asistenciales (médicos de familia, especialistas y psiquiatras), así como una educación clara al paciente, para evitar duplicidades o combinaciones peligrosas de medicamentos (Sempere & Jurado, 2017).

En el ámbito de las urgencias hospitalarias, relacionadas con estudios sobre ingresos y reingresos provocados por RAM en personas ancianas, se siguen identificando múltiples factores de riesgo relacionados con medicamentos. Entre ellos se incluyen la polifarmacia, incumplimiento terapéutico, deterioro cognitivo, duración de la estancia hospitalaria, enfermedad renal y el uso de medicamentos de alto riesgo. Estos factores aumentan la vulnerabilidad a padecer una RAM, lo que a su vez incrementa las tasas de hospitalización. Las intervenciones para reducir estas hospitalizaciones señalan que la participación de farmacéuticos y los programas de formación a distintos niveles han demostrado resultados prometedores. Además, el uso de sistemas de soporte para la toma de decisiones clínicas puede ayudar en la identificación y prevención de RAM (Linkens et al., 2020).

Sin embargo, los autores destacan que existe una considerable heterogeneidad en las definiciones y metodologías utilizadas en los estudios revisados, lo que dificulta la comparación y generalización de los resultados. Por lo tanto, se enfatiza la necesidad de establecer definiciones estandarizadas y realizar investigaciones adicionales para desarrollar estrategias efectivas que reduzcan las hospitalizaciones relacionadas con medicamentos en la población de adultos de edad avanzada (Linkens et al., 2020).

En el contexto de un hospital de tamaño medio con recursos limitados, la utilización y el manejo de datos sobre ingresos hospitalarios causados por RAM en pacientes polimedicados de edad avanzada representa un desafío significativo. La polimedicación incrementa el riesgo de interacciones medicamentosas y eventos adversos, lo que a su vez puede derivar en hospitalizaciones inevitables. Además, la recopilación, análisis y utilización efectiva de estos datos suele verse obstaculizada por la falta de sistemas informatizados robustos, la escasez de personal especializado y la fragmentación de los registros clínicos.

En entornos con recursos limitados, la recolección de datos sobre RAM en pacientes polimedicados a menudo se realiza mediante sistemas pasivos, como notificaciones espontáneas que dependen de la iniciativa del personal médico, por lo que estos métodos son propensos a la subnotificación por requerir tiempo y recursos en entornos con una alta carga laboral, o bien, por falta de formación en farmacovigilancia, lo que dificulta la identificación y notificación de las RAM. Como alternativa, algunos hospitales implementan sistemas activos de farmacovigilancia, donde equipos multidisciplinarios revisan historias para identificar posibles RAM (Alhawassi et al., 2014). No obstante, la falta de integración entre sistemas electrónicos dificulta la consolidación de la información.

A pesar de los avances tecnológicos, muchos hospitales con recursos limitados enfrentan barreras significativas en el manejo de datos sobre RAM en pacientes polimedicados. La adopción de sistemas electrónicos interoperables, junto con estrategias de farmacovigilancia activa, podría mejorar la detección y prevención de estos eventos. Sin embargo, se requiere mayor inversión en infraestructura y capacitación para optimizar estos procesos.

En el contexto de un hospital de tamaño medio, la UF desempeña un papel clave en el estudio y monitorización de los ingresos hospitalarios provocados por RAM. Su labor permite identificar los medicamentos más frecuentemente implicados en estas reacciones, analizar los factores de riesgo asociados y desarrollar estrategias para minimizar su impacto.

El análisis de los ingresos hospitalarios por RAM es una tarea compleja que requiere de la integración y procesamiento de grandes volúmenes de datos clínicos que plantea una serie de brechas o desafíos:

1. Acceso y procesamiento de datos clínicos

- La información sobre los episodios de hospitalización, diagnósticos y medicamentos implicados se encuentra dispersa en diferentes sistemas (historias clínicas, prescripción electrónica, laboratorio, farmacia, etc.) por lo que el acceso y procesamiento de datos clínicos es una tarea compleja y costosa (Nebeker et al., 2004).
- La fragmentación de los datos o silos de información dificultan la trazabilidad entre medicamentos prescritos y eventos adversos observados.
- Las restricciones legales y éticas limitan el uso de datos clínicos reales para análisis exploratorios.
- Es necesario garantizar la correcta extracción, transformación y carga (ETL) de los datos provenientes del sistema de gestión del hospital (Bates et al., 2003).
- La ausencia de mecanismos efectivos de anonimización frena el desarrollo de herramientas analíticas en entornos de prueba o preproducción.

2. Fiabilidad y calidad de los datos

- La detección de RAM depende de la precisión de los diagnósticos y registros médicos, por lo que la existencia de datos incompletos o errores en la codificación puede afectar a la validez de los análisis (Pirmohamed et al., 2004).
- La presencia de registros incompletos e inconsistencias en la codificación de diagnósticos o fármacos redunda en una baja calidad de los datos.

3. Interpretación de los resultados

- Identificar si un ingreso hospitalario es consecuencia directa de una RAM requiere un análisis riguroso, debiéndose tener en cuenta factores como interacciones medicamentosas, condiciones preexistentes y errores de prescripción, pudiendo influir en la ocurrencia de RAM (Howard et al., 2007).
- La identificación de patrones complejos como interacciones medicamentosas, requiere de una explotación de consultas analíticas avanzadas o modelos predictivos.

4. Repercusión en la seguridad del paciente

- Comprender los patrones de RAM permite diseñar estrategias preventivas, como la mejora en la prescripción o la implementación de alertas en los sistemas clínicos (Weltgesundheitsorganisation & Collaborating Centre for International Drug Monitoring, 2002).
- La farmacovigilancia hospitalaria puede contribuir a la optimización del uso de medicamentos, reduciendo efectos adversos y mejorando la calidad asistencial (Ordoñez et al., 2023).
- Para validar algoritmos, entrenar modelos, detectar patrones en reacciones adversas o hacer pruebas sin comprometer la privacidad de los pacientes, son necesarios repositorios anonimizados para experimentar o simular interacciones entre medicamentos, comprender mejor los riesgos y beneficios de la combinación de medicamentos, garantizar que los pacientes reciban los tratamientos más adecuados y acelerar el desarrollo de soluciones tecnológicas en entornos seguros de pruebas.

Dado el impacto clínico y económico de las RAM en la hospitalización, este trabajo pretende contribuir a la mejora de la seguridad del paciente mediante el análisis de datos almacenados en un repositorio externo anonimizado, mediante el cual la UF podrá analizar y detectar patrones de RAM relacionados con ingresos hospitalarios, revisar los medicamentos implicados en dichos ingresos, recopilar información sobre la polimedicación y generar información valiosa empleando las tecnologías de la información para la toma de decisiones en la práctica clínica (Bermeo Cerón et al., 2024).

2.2. Estado del arte

Las RAMs constituyen un problema de salud pública significativo, pueden contribuir a una proporción considerable de ingresos hospitalarios, prolongaciones en la estancia de los pacientes y en consecuencia al aumento de los costes sanitarios (Howard et al., 2007). Diversos estudios han demostrado que una parte importante de estas hospitalizaciones podrían evitarse mediante una mejor identificación y gestión de las RAMs (Bates et al., 2003).

Como señalan estudios recientes, los sistemas actuales de farmacovigilancia, aunque esenciales, presentan limitaciones críticas en la identificación de RAM asociadas a

polimedicación en pacientes de edad avanzada. La mayoría de los sistemas se basan en notificaciones de señales individuales de RAM, ignorando aquellas derivadas de combinaciones complejas de fármacos, lo que dificulta la identificación de RAMs graves (Hohl et al., 2013).

Los sistemas de notificación espontánea de RAM presentan un importante sesgo de subregistro, ya que solamente capturan entre el 5% y 10% de los casos reales, porcentaje que es aún menor en pacientes polimedicados de edad avanzada debido a que con frecuencia se atribuyen erróneamente sus síntomas a fragilidad o comorbilidades subyacentes.

Los sistemas actuales de farmacovigilancia carecen de datos longitudinales y contextuales clave para ancianos, como deterioro renal/hepático, polifarmacia o adherencia (Zazzara et al., 2021b), lo que limita la detección de interacciones medicamentosas evitables. Existen estudios que revelan que el 30% de las hospitalizaciones por RAM en mayores están ligadas a estas interacciones, pero las herramientas disponibles no las priorizan debido a su diseño centrado en señales simples, no en complejidad geriátrica.

La literatura propone innovaciones para mejorar la farmacovigilancia en polimedicación geriátrica, destaca una brecha tecnológica en la detección de RAM complejas que podría abordarse mediante un repositorio externo con datos anonimizados empleando modelos predictivos como machine learning (ML) o inteligencia artificial (IA) y aplicando algoritmos para identificar interacciones entre combinaciones de fármacos e incorporando variables geriátricas en análisis longitudinales.

En un hospital de tamaño medio donde se trabaja con datos sensibles, como es la información clínica, no es recomendable realizar estudios de farmacovigilancia directamente en sistemas de producción (ej.: historias clínicas electrónicas, diagnósticos y prescripciones en tiempo real) sin un marco seguro y controlado debido a riesgos éticos, legales y técnicos. Sin embargo, con las salvaguardas adecuadas, pueden extraerse datos y aplicar técnicas de anonimización para realizar análisis retrospectivos.

En este contexto, es necesario un proceso ETL para extraer, transformar y cargar los datos en un repositorio externo sobre el que se realizarán los análisis sobre RAM. Si los **datos** deben estar **seudonimizados** en destino para proteger la privacidad de los pacientes, el proceso ETL extraerá los datos de los sistemas del hospital, los transformará sustituyendo los

identificadores personales (nombres, DNI o números de paciente) por códigos o seudónimos consistentes y los cargará en el repositorio externo. Lo característico de esta técnica es la existencia de una clave de correspondencia guardada en un lugar seguro que permite, si es necesario, volver a identificar al paciente. De ese modo, los datos seguirán siendo individualizables, pero protegidos (*Aproximación a los espacios de datos desde la perspectiva del RGPD | AEPD*, 2023).

Por otra parte, si los **datos** deben ser **anonimizados**, el proceso ETL va más allá. No solo eliminará o codificará los identificadores directos, sino que también transformará atributos indirectos (como género, edad, fechas, dosis de medicamentos, etc.), que combinados, podrían revelar la identidad de una persona. Esta transformación puede implicar generalizar valores (por ejemplo, convertir las edades exactas en rangos de edades) o incluso eliminar registros para evitar la reidentificación. A diferencia de la seudonimización, no existe forma de volver a saber quién era cada paciente. Una vez los datos se cargan en el repositorio externo, la información personal habrá desaparecido de manera irreversible (*Aproximación a los espacios de datos desde la perspectiva del RGPD | AEPD*, 2023).

En resumen, un ETL con seudonimización se diseñará para preservar el vínculo con la identidad real bajo control, mientras que un ETL con anonimización se enfoca en eliminar cualquier posibilidad de identificación, aunque este segundo caso puede implicar cierta pérdida de precisión en los datos.

2.2.1. Sistemas orientados a la notificación de sospechas por RAM

NotificaRAM

Es el sistema oficial de notificación electrónica de sospechas de RAM en España, gestionado por la Agencia Española de Medicamentos y Productos Sanitarios (AEMPS). Su objetivo es facilitar la comunicación directa de posibles efectos adversos tanto por parte de profesionales sanitarios como de ciudadanos, contribuyendo así a la seguridad del uso de medicamentos (*AEMPS*, 2015).

NotificaRAM es un sistema de notificación espontánea, considerado el método más eficiente para identificar nuevos riesgos asociados a medicamentos, especialmente aquellos poco frecuentes o graves. Este sistema complementa al tradicional método de la tarjeta amarilla,

permitiendo una notificación más ágil y accesible a través de formularios electrónicos disponibles en www.notificaram.es. Los formularios están adaptados tanto para profesionales sanitarios como para ciudadanos, y en caso de que una Comunidad Autónoma disponga de su propio formulario electrónico, el sistema redirige automáticamente al usuario al formulario correspondiente (*AEMPS*, 2015).

Las notificaciones recibidas a través de NotificaRAM son evaluadas y registradas por el Sistema Español de Farmacovigilancia de Medicamentos de Uso Humano (SEFV-H), una red coordinada por la AEMPS y constituida por los Centros Autonómicos de Farmacovigilancia. Estos centros son responsables de analizar las notificaciones y registrarlas en la base de datos FEDRA (Farmacovigilancia Española, Datos de Reacciones Adversas), que recopila información anonimizada sobre las RAM notificadas en todo el país (*AEMPS*, 2015).

La información recopilada en FEDRA es fundamental para detectar señales de nuevos riesgos asociados a medicamentos, lo que permite a la AEMPS, en coordinación con otras agencias europeas, tomar decisiones regulatorias para garantizar que los beneficios de los medicamentos superen sus posibles riesgos (*AEMPS*, 2015).

EudraVigilance

Es la plataforma oficial de farmacovigilancia de la Unión Europea, desarrollada y gestionada por la Agencia Europea de Medicamentos (EMA). Su propósito principal es recopilar, gestionar y analizar informes de sospechas de reacciones adversas a medicamentos (RAM) en los países del Espacio Económico Europeo (EEE), contribuyendo así a la detección temprana de señales de seguridad y a la protección de la salud pública.

Es un sistema de notificación electrónica centralizada que permite la transmisión segura y estandarizada de informes de casos individuales de seguridad (ICSRs, por sus siglas en inglés). Estos informes pueden ser enviados por titulares de autorizaciones de comercialización, autoridades nacionales competentes y patrocinadores de ensayos clínicos. El sistema facilita la recolección de datos tanto de medicamentos autorizados como de aquellos en fase de investigación clínica (*EMA*, 2025).

La EMA, en colaboración con las autoridades nacionales competentes de los Estados miembros del EEE, es responsable de la evaluación y registro de las sospechas de RAM

notificadas a través de EudraVigilance. Dentro de la EMA, el Comité de Evaluación de Riesgos de Farmacovigilancia (PRAC) desempeña un papel clave en la revisión de los datos para identificar señales de seguridad y recomendar acciones regulatorias cuando sea necesario (*EMA*, 2025).

EudraVigilance también interactúa con otras bases de datos internacionales y contribuye a la base de datos pública de informes de sospechas de reacciones adversas a medicamentos, accesible a través del portal www.adrreports.eu, promoviendo así la transparencia y el acceso a la información sobre la seguridad de los medicamentos (*Base de datos europea de informes de presuntas reacciones adversas*, s.f.).

FAERS

La FDA Adverse Event Reporting System (FAERS) es la plataforma oficial de farmacovigilancia de la Administración de Alimentos y Medicamentos de los Estados Unidos (FDA), diseñada para recopilar y analizar informes de sospechas de RAM, errores de medicación y quejas sobre la calidad de productos farmacéuticos y biológicos terapéuticos. Su objetivo principal es respaldar la vigilancia de seguridad postcomercialización de estos productos, permitiendo la identificación temprana de posibles riesgos y la implementación de medidas regulatorias para proteger la salud pública.

FAERS opera como un sistema de notificación espontánea, donde los informes pueden ser presentados voluntariamente por profesionales de la salud, consumidores y otras partes interesadas. Además, los fabricantes de medicamentos están obligados a reportar a la FDA cualquier evento adverso que reciban. La estructura informática de FAERS cumple con las directrices internacionales de reporte de seguridad emitidas por la Conferencia Internacional sobre Armonización (ICH E2B), asegurando la estandarización y calidad de los datos recopilados (Sakaeda et al., 2013).

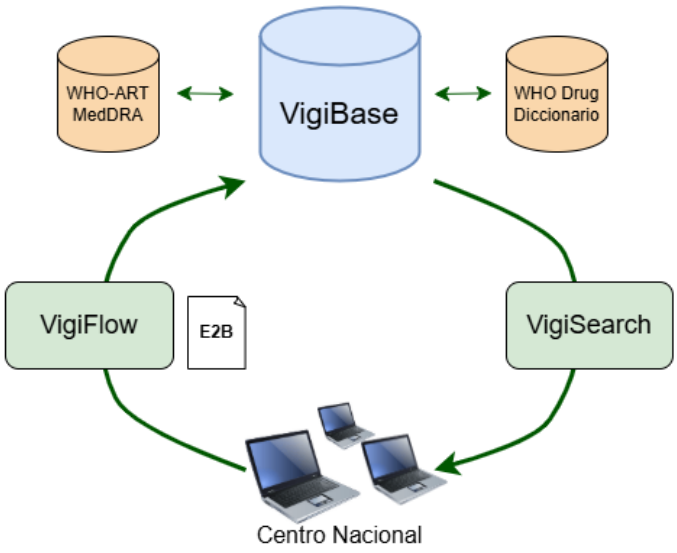
VigiBase

Es la base de datos global de farmacovigilancia de la Organización Mundial de la Salud (OMS), gestionada por el Uppsala Monitoring Centre (UMC) desde 1968. Su objetivo principal es

recopilar, gestionar y analizar informes de sospechas de RAM provenientes de los países miembros del Programa de la OMS para la Vigilancia Internacional de Medicamentos (WHO PIDM) (Uppsala Monitoring Centre, 2025).

VigiBase funciona como un sistema de notificación espontánea y es la mayor base de datos de informes de seguridad de medicamentos a nivel mundial, con más de 5 millones de reportes anónimos de sospechas de reacciones adversos. Los informes, conocidos como Informes Individuales de Seguridad de Casos (ICSRs), son enviados a VigiBase desde los centros nacionales de farmacovigilancia en un formato estandarizado internacional (ICH E2B). Estos informes incluyen datos sobre medicamentos, vacunas, productos biológicos y terapias tradicionales, y se codifican utilizando sistemas estandarizados como MedDRA (reacciones adversas) y WHODrug (medicamentos) para facilitar su análisis (Uppsala Monitoring Centre, 2025).

Figura 1. Flujo de datos hacia VigiBase y consulta desde VigiSearch.



Fuente: Elaboración propia, 2025.

Tabla 2. Principales herramientas de notificación de sospechas de RAM.

Herramienta	Función principal	Gestora	Base de datos
NotificaRAM	Recogida de sospechas de RAM en territorio español	Agencia Española de Medicamentos y Productos Sanitarios (AEMPS)	Sistema Español de Farmacovigilancia (SEFV-H)

EudraVigilance	Gestión centralizada de RAM en Europa	Agencia Europea de Medicamentos (EMA)	EudraVigilance
FAERS	Recogida de informes voluntarios y obligatorios de RAM en EE.UU.	U.S. Food and Drug Administration (FDA)	FDA Adverse Event Reporting System (FAERS)
VigiBase	Plataforma global de RAM basada en reportes de centros nacionales de vigilancia	Uppsala Monitoring Centre (por delegación de la OMS)	VigiBase (Base global de la OMS)

Fuente: Elaboración propia, 2025.

2.2.2. Sistemas orientados al análisis de datos de farmacovigilancia

OpenVigil

Es un sistema de código abierto diseñado para el análisis de datos de farmacovigilancia, especialmente útiles en investigación y detección sobre la seguridad de los medicamentos. Su objetivo principal es facilitar la detección de señales de seguridad y la identificación de patrones en los datos de RAM (Böhm et al., 2016).

OpenVigil permite extraer y analizar datos de bases de datos públicas de RAM como FAERS (FDA, EE.UU.) y VigiBase (OMS). Identifica posibles asociaciones entre medicamentos y RAM empleando métodos de análisis de desproporcionalidad, como el cálculo del Proportional Reporting Ratio (PRR). Estos métodos permiten detectar señales de seguridad al comparar la frecuencia de una RAM específica con la frecuencia esperada en la base de datos (Böhm et al., 2016).

La herramienta ofrece capacidades de visualización que incluyen la presentación de resultados en tablas interactivas, gráficos y la posibilidad de exportar los datos para su análisis en software estadístico. Estas visualizaciones ayudan a los usuarios a interpretar los datos y a identificar patrones o tendencias en las RAM (Böhm et al., 2016).

A pesar de ser una solución diseñada para analizar datos de farmacovigilancia pública, no puede conectarse directamente a repositorios internos de hospitales sin realizar una adaptación. Por otra parte, está destinada al análisis de datos agregados, no para registros

clínicos individuales, por lo que es un sistema limitado para la evaluación detallada de causalidad paciente a paciente.

REDCap

REDCap (Research Electronic Data Capture) es una plataforma web desarrollada por la Universidad de Vanderbilt en Estados Unidos. Su creación respondió a la necesidad institucional de contar con una herramienta segura y conforme a las regulaciones de privacidad, diseñada para la recopilación y gestión de datos en investigaciones clínicas y estudios observacionales (Harris et al., 2009).

La plataforma permite la implementación de formularios estandarizados para la captura de datos sobre RAM, lo que facilita la recopilación sistemática y coherente de información. Aunque REDCap no realiza análisis automatizados de desproporcionalidad, los datos pueden exportarse a herramientas estadísticas externas para realizar análisis avanzados. Ofrece funcionalidades básicas de visualización, como gráficos y tablas resumen, que permiten a los usuarios obtener una visión general de los datos recopilados (Harris et al., 2009).

Aunque REDCap no ofrece un módulo de farmacovigilancia estándar, su flexibilidad y capacidad modular permitiría desarrollar soluciones externas adaptadas para la gestión y análisis de RAM en entornos clínicos. Por ejemplo, empleando el módulo Clinical Data Interoperability Services (CDIS), REDCap puede conectar con historiales clínicos e importar datos estructurados de registros electrónicos de salud mediante comunicaciones protegidas utilizando el estándar FHIR que permite compartir información de diferentes sistemas de salud de forma segura y eficiente. Esto facilita la recopilación de información clínica relevante para la farmacovigilancia. (Harris et al., 2009).

PV-Works

Es una solución comercial para gestionar de manera integral la farmacovigilancia en entornos humanos y veterinarios, permitiendo recopilar, analizar y reportar datos relacionados con RAM en una única plataforma, tanto en contextos clínicos como poscomercialización («Ennov Pharmacovigilance Suite», s.f.).

Su diseño modular facilita la adaptación a las necesidades específicas de cada organización, permitiendo la implementación en instalaciones locales o en la nube. Además, es compatible con diversos productos médicos, incluyendo medicamentos tradicionales, productos biológicos y dispositivos médicos.

PV-Works incorpora herramientas avanzadas para la detección y gestión de señales de seguridad. Utiliza métodos estadísticos como el Proportional Reporting Ratio (PRR), el Reporting Odds Ratio (ROR) y el Multi-item Gamma Poisson Shrinker (MGPS) para identificar posibles asociaciones entre medicamentos y reacciones adversas. Estas técnicas permiten analizar grandes volúmenes de datos y detectar señales que podrían indicar riesgos emergentes para la salud pública.

La plataforma ofrece capacidades robustas de visualización como la generación de informes personalizados y la exportación de datos para análisis adicionales en otras herramientas.

Aunque es una solución comercial muy potente, su implementación en hospitales medianos puede ser muy costosa y requiere de un período de adaptación. Esta herramienta está más pensada para para industrias farmacéuticas que para la gestión interna hospitalaria.

VigiMine

Es una herramienta avanzada de minería de datos diseñada y gestionada por el Uppsala Monitoring Centre (UMC), organismo colaborador de la Organización Mundial de la Salud (OMS) para la farmacovigilancia. Su principal función es identificar patrones inusuales en la base de datos mundial de sospechas RAM, conocida como VigiBase.

A diferencia de otras herramientas tradicionales de consulta, VigiMine permite un análisis proactivo y automatizado de los millones de reportes almacenados en VigiBase, facilitando la detección temprana de señales de seguridad. Estas señales pueden ser indicios de nuevas RAM previamente no reconocidas o de cambios en la frecuencia o gravedad de efectos ya conocidos.

El sistema se basa en métodos estadísticos bayesianos, especialmente el Information Component (IC), una técnica que mide la desproporcionalidad en la aparición de un evento adverso asociado a un medicamento específico en comparación con la frecuencia esperada.

Este método ayuda a reducir el ruido de fondo y a priorizar asociaciones significativas que requieran un análisis más profundo por parte de los centros nacionales de farmacovigilancia.

En cuanto a sus capacidades de visualización de datos, VigiMine ofrece una visualización interactiva y jerarquizada de los datos, presentando las combinaciones de medicamento-evento más relevantes mediante rankings, filtros dinámicos y alertas por regiones o grupos poblacionales. Esto permite a los expertos explorar fácilmente posibles problemas de seguridad emergentes y compartir hallazgos con las autoridades regulatorias y la comunidad médica internacional.

VigiMine representa una herramienta crítica dentro del ecosistema de farmacovigilancia global. Potencia la capacidad de los sistemas sanitarios para proteger a los pacientes, ofreciendo una plataforma eficiente, basada en inteligencia estadística, para detectar de forma temprana las amenazas potenciales asociadas a los medicamentos en uso (*Uppsala Reports January 2010.pdf*, s.f.).

Aunque la herramienta facilita la detección temprana de señales de seguridad, lo hace centrándose en señales poblaciones accediendo a datos agregados de VigiBase y no trabaja directamente con repositorios locales, por lo que no permite realizar análisis profundos sobre cohortes concretas o historiales clínicos individuales de pacientes.

VigiRank

Es un sistema de detección de señales desarrollado por el Uppsala Monitoring Centre (UMC) que se basa en un modelo multivariado de regresión logística con regularización (shrinkage logistic regression) para evaluar la probabilidad de que una combinación medicamento-RAM represente una señal de seguridad real, reflejando una puntuación para cada par medicamento-RAM. Las combinaciones se ordenan según esta puntuación para priorizar su evaluación clínica (Caster et al., 2014).

VigiRank puede identificar patrones sobre RAM considerando diversas variables que capturan diferentes aspectos de la solidez de la evidencia en los informes de RAM, incluyendo:

- El número de informes recibidos en los últimos años.

- La desproporcionalidad o presencia de una frecuencia inusualmente alta de la combinación medicamento-RAM en comparación con otras.
- La disponibilidad de descripciones detalladas de los casos.
- La diversidad geográfica de los informes en los países de origen.

En cuanto a sus capacidades de visualización de datos, VigiRank se integra en herramientas de análisis como VigiLyze, proporcionando gráficos de tendencias temporales, mapas de distribución geográfica y tablas de puntuaciones que permiten comparar y priorizar combinaciones medicamento-RAM (Caster et al., 2014).

A pesar de que alguna de sus funcionalidades podría responder a algún requisito de este proyecto, este sistema está diseñado para priorizar la detección de señales de RAM a partir de grandes bases de datos como VigiBase empleando métodos de "ranking", por lo que su implementación se vería limitada para analizar datos clínicos hospitalarios específicos, no permitiendo trabajar con variables clínicas detalladas necesarias para realizar estudios retrospectivos personalizados.

Tabla 3. Comparación de herramientas para análisis de RAM.

Herramienta	Función principal	Métodos clave	Visualización de datos
OpenVigil	Minería de datos de farmacovigilancia basada en la base de datos FAERS (FDA).	PRR (Proportional Reporting Ratio), ROR (Reporting Odds Ratio), IC.	Gráficos de líneas y barras, tablas de frecuencia, filtrado interactivo de criterios.
REDCap	Recolección estructurada y gestión de datos clínicos, incluidas las notificaciones de RAM	Estadística descriptiva (tablas, gráficos, medidas de tendencia o dispersión), análisis temporal, integración con R y SAS	Dashboards personalizados, gráficas de líneas, exportación a plataformas estadísticas externas
PV-Works	Gestión integral de farmacovigilancia regulatoria y comercial.	Modelos de detección por agrupamiento, evaluaciones de causalidad.	Paneles interactivos, alertas gráficas, mapas de calor.

VigiMine	Detección automática de señales en la base global VigiBase.	Modelos bayesianos (IC025), estadística de desproporcionalidad.	Gráficos de señales, análisis por frecuencia/tiempo, dashboards interactivos.
VigiRank	Priorización de señales RAM considerando múltiples dimensiones.	Modelos multivariados.	Rangos de prioridad, mapas de calor, listas ordenadas de riesgo.

Fuente: Elaboración propia, 2025.

2.2.3. Técnicas de anonimización

En esta sección se proporciona una visión general de las técnicas de anonimización fundamentales para proteger la privacidad de los pacientes mientras se permite el uso de datos para investigación y mejora de la gestión sanitaria. Estas técnicas son esenciales para cumplir normativas de privacidad del Reglamento General de Protección de Datos (RGPD) en Europa (Delgado, 2024).

Conviene distinguir entre anonimización y seudonimización, dos conceptos que se suelen confundir en ocasiones. Su diferencia radica en las garantías y en el grado de protección de los derechos de los interesados. Por una parte, los datos anonimizados no guardan relación con una persona física identificada o identificable (Considerando 26 del RGPD) y no están bajo el ámbito de aplicación del RGPD, en cambio, los datos seudonimizados y la información adicional que se pueda vincular con dicho conjunto de datos sí que lo están, por lo que puede estar sujeta al cumplimiento de medidas técnicas y organizativas destinadas para garantizar que los datos no puedan ser atribuidos a una persona física (Vollmer, 2023).

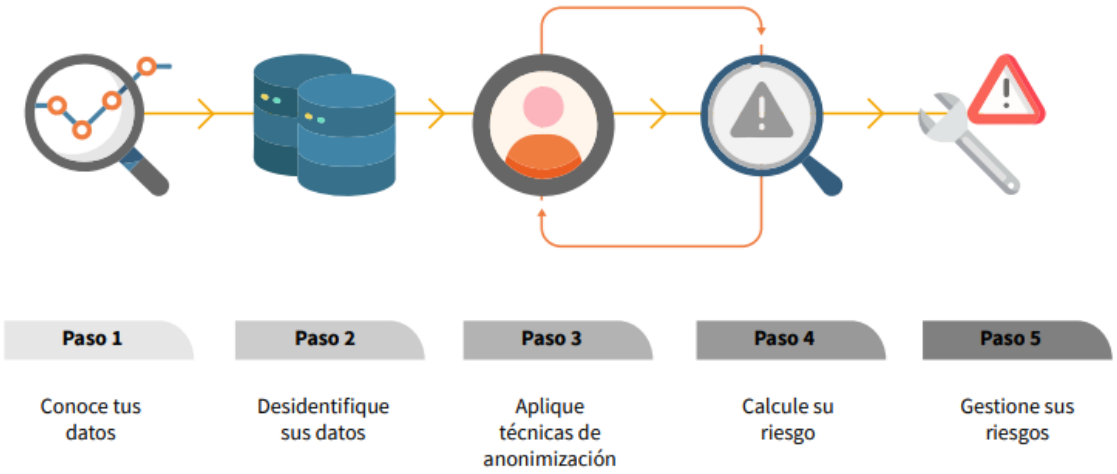
La transformación de un conjunto de datos personales en información anonimizada o seudonimizada requiere de un tratamiento sobre dichos datos, y así como la técnica de anonimización genera un único y nuevo conjunto de datos, la seudonimización genera dos nuevos conjuntos de datos, por una parte, la información seudonimizada y, por otra parte, la información adicional necesaria para revertir la seudonimización empleando una clave de correspondencia y de ese modo poder reidentificar al paciente. Mientras que sobre el conjunto de datos anonimizados, desde el punto de vista del RGPD, se debe garantizar la robustez del proceso de anonimización contra la posible reidentificación y no requiere

garantías adicionales, para el conjunto de datos seudonimizados se debe impedir la reidentificación sin disponer de la información adicional, limitar el tratamiento (finalidades, periodo de conservación, comunicación de datos seudonimizados) y por último, ofrecer garantías contra brechas de seguridad, tanto sobre el conjunto seudonimizado como de la información adicional que se pueda vincular con dicho conjunto de datos (*Anonimización y seudonimización | AEPD*, 2021).

La anonimización es crucial por varias razones:

- Garantiza que los datos puedan ser utilizados sin comprometer la identidad de los pacientes.
- Permite a las instituciones sanitarias cumplir con normativas y leyes que exigen la protección de datos personales sensibles.
- Los datos anonimizados pueden ser compartidos de forma segura entre instituciones para acelerar el progreso de la investigación y el desarrollo de medicamentos y tratamientos.
- El proceso de anonimización elimina los datos sensibles.

Figura 2. Proceso de Anonimización.



*Fuente: (*Guía y Herramienta básica de anonimización | AEPD*, 2022).*

A pesar de los beneficios de la anonimización, presenta una serie de desafíos:

- Para garantizar la irreversibilidad en el diseño es fundamental adoptar métodos robustos, revisarlos y actualizarlos con cierta periodicidad.
- Encontrar un equilibrio entre privacidad y funcionalidad es un desafío constante.
- La implementación de sistemas de anonimización requiere inversión en tecnología y capacitación, lo que puede implicar un aumento de costes y recursos para las organizaciones.
- Las regulaciones y estándares para aplicar la anonimización puede variar entre países, por lo que puede complicar su implementación en organizaciones globales.

A continuación, pasamos a describir los métodos clásicos de anonimización utilizados en el ámbito de la salud:

Supresión (Eliminación Directa).

Consiste en eliminar identificadores directos (nombre, DNI, dirección, teléfono) de los datos. La limitación no siempre garantiza el anonimato, ya que combinando otros datos (edad, género, diagnóstico) se podría reidentificar al paciente. Este método pretende eliminar la "contaminación" (valores poco usuales o fuera de rango) para evitar un aumento de la probabilidad de reidentificación.

Generalización.

Consiste en reducir la precisión de los datos para hacerlos menos identificables, como la creación de rangos en caso de atributos numéricos o el establecimiento de jerarquías para los atributos nominales. De este modo, podemos incrementar el número de registros que poseen los mismos valores para un conjunto de atributos cuasi-identificadores con el objeto de satisfacer los requisitos de privacidad a la vez que cumplimos con la finalidad del tratamiento de los datos.

Ejemplo:

Cambio de la Edad exacta por Rango de edad (ej. "30-39 años").

Cambio de la Fecha exacta por mes, trimestre o año (ej. "2020" en lugar de "15/03/2020").

Seudonimización.

Consiste en sustituir identificadores directos por códigos o alias (seudónimos), pero tiene el inconveniente de que no es una anonimización total porque permite la reidentificación con una clave.

Ejemplo: Reemplazar el nombre del paciente por un código "PAC-123".

Agregación de datos.

Consiste en presentar los datos en forma de resúmenes estadísticos (medias, conteos, porcentajes) en lugar de registros individuales. Tiene la ventaja de que reduce el riesgo de reidentificación y el inconveniente de que pierde detalle para análisis individualizados.

Perturbación de datos.

Consiste en añadir "ruido" a los datos para evitar la identificación. Es utilizado en bases de datos con variables numéricas sensibles por lo que puede afectar a la precisión de los análisis.

Ejemplo: Modificar ligeramente la edad (ej. ±2 años) o la fecha de ingreso.

Intercambio de valores (Shuffling o Permutación)

Consiste en reordenar valores entre registros para desvincular atributos identificables con el objetivo de romper relaciones entre datos sin alterar estadísticas globales.

Ejemplo: Intercambiar códigos postales entre pacientes de la misma cohorte.

K-anonimidad (Modelo clásico de privacidad)

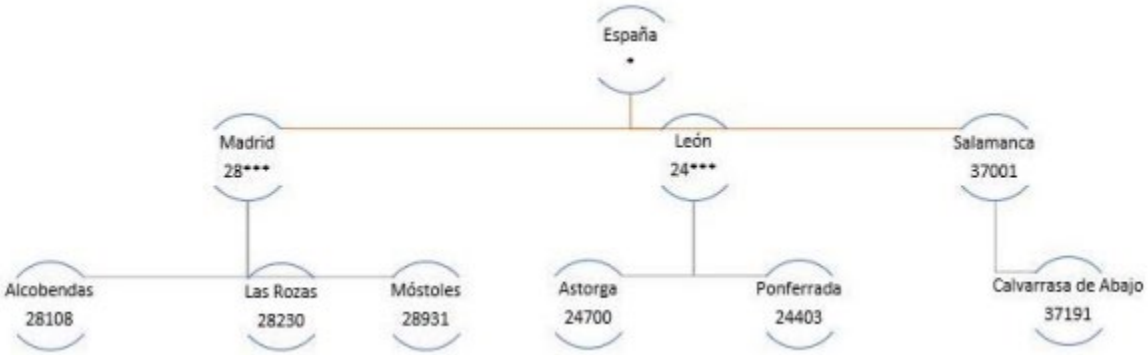
Es una medida del riesgo de que agentes externos puedan obtener información de carácter personal a partir de datos anonimizados y consiste en garantizar que cada registro en un conjunto de datos sea indistinguible de al menos k-1 registros respecto a ciertos atributos cuasi-identificadores (aquellos que de forma aislada no identifican a un individuo, pero agrupados con otros atributos pueden señalar de forma unívoca a un sujeto). De este modo,

la probabilidad de identificar a un individuo concreto en base a un conjunto de cuasi-identificadores es como máximo 1/K, por lo que para conseguir un bajo riesgo de reidentificación debe garantizarse un valor mínimo de K cuando se pretenda llevar a cabo el diseño de un proceso de anonimización o disociación de datos (*Anonimización y seudonimización | AEPD*, 2021). Respecto al valor de K se debe conseguir un equilibrio, porque a mayor valor de K se puede perder fidelidad en los datos origen, por lo que deberemos determinar en qué grado esa pérdida de fidelidad puede influir en la finalidad del tratamiento por la pérdida de información relevante.

Para implementar la K-anonimización y no introducir perturbación en los datos se suelen combinar métodos como la generalización y la supresión, sin perder de vista el objetivo de proteger la privacidad sin destruir el valor analítico del conjunto de datos.

En la siguiente figura se muestra un ejemplo de una clasificación de jerarquía para el atributo "Código postal" y a continuación se muestran dos tablas, una primera tabla con los datos originales, y una segunda tabla como resultado de aplicar una generalización del atributo "Edad" (dentro de un rango numérico), así como la supresión de datos poco usuales o fuera de rango que pueden no formar parte del estudio que estamos realizando. Además, al actuar sobre dos atributos cuasi-identificadores y aplicar los dos métodos, conseguimos que la tabla resultante sea 2-anónima con K=2.

Figura 3. Jerarquía para el campo Código Postal.



*Fuente: (*Anonimización y seudonimización | AEPD*, 2021).*

Tabla 4. Tabla original y tabla 2-anónima (Generalización + Eliminación directa).

Código postal	Edad	Colesterol
37003	40	S
28108	44	S
24700	37	N
24700	37	N
37003	44	S
28108	40	S
37891	33	N
50011	13	S

Tabla original

Código postal	Edad	Colesterol
37***	40 - 49	S
28***	40 - 49	S
24700	30 - 39	N
24700	30 - 39	N
37***	40 - 49	S
28***	40 - 49	S
37***	30 - 39	N

Tabla 2-anónima

Fuente: (*Anonimización y seudonimización | AEPD*, 2021)

Cuando se intenta anonimizar un conjunto de datos, se recurre a técnicas como la generalización y la eliminación, pero ambas introducen distorsiones. La eliminación de datos puede afectar gravemente la representatividad del conjunto, mientras que la generalización reduce el valor informativo al hacer que los datos pierdan precisión. Si bien estas distorsiones son más graves en conjuntos pequeños, en grandes volúmenes de datos, eliminar unos pocos valores puede ser preferible a aplicar generalizaciones muy amplias. Además, lograr una anonimización que cumpla con el criterio de k-anonimato implica resolver un problema matemáticamente complejo, catalogado como NP-duro, un problema para el que no se conoce una solución eficiente y que tiene una difícil solución desde el punto de vista computacional (*Anonimización y seudonimización | AEPD*, 2021).

Existen diversas herramientas que implementan algoritmos para facilitar el proceso de k-anonimización:

ARX Data Anonymization Tool

Es una herramienta de código abierto que permite aplicar técnicas de privacidad como la k-anonimidad sobre grandes conjuntos de datos, así como modelos de transformación como el muestreo aleatorio o la microagregación. Cuenta con una interfaz gráfica multiplataforma y

de una API de integración con Java para implementar la anonimización de datos desde software desarrollado en este mismo lenguaje de programación (*Anonimización y seudonimización | AEPD*, 2021).

Enlace de descarga: https://arx.deidentifier.org/downloads/

Herramienta de anonimización UTD

Es una herramienta de código abierto desarrollada por UT Dallas Data Security y Privacy Lab, que implementa métodos de anonimización que pueden ser aplicados sobre un conjunto de datos, o bien, a través de librerías implementadas dentro de otras aplicaciones (*Anonimización y seudonimización | AEPD*, 2021).

Enlace de descarga: http://cs.utdallas.edu/dspl/cgibin/toolbox/index.php?go=download

Amnesia

Es una herramienta de código abierto desarrollada por OpenAIRE que utiliza un backend de Java para la anonimización de datos que funciona tanto en Windows con en Linux (*Anonimización y seudonimización | AEPD*, 2021).

Ofrece dos modalidades de uso:

1. **Versión en línea:** Disponible para pruebas con conjuntos de datos pequeños, accesible a través de su sitio web oficial (https://amnesia.openaire.eu/amnesia/).
2. **Aplicación local**: Permite ejecutar Amnesia en servidores propios para garantizar que los datos sensibles no salgan de las instalaciones de la institución (https://github.com/dTsitsigkos/Amnesia).

Consideraciones importantes:

- Se debe encontrar un equilibrio entre utilidad y privacidad, teniendo en cuenta que algunos métodos pueden llegar a degradar excesivamente la calidad de los datos.
- En el contexto legal, normativas como RGPD (UE) o HIPAA (EEUU) establecen requisitos para la anonimización que deben tenerse en cuenta.

- Los métodos clásicos en ocasiones pueden llegar a ser vulnerables a ataques de reidentificación con datos auxiliares, por lo que puede llegar a existir un riesgo residual de vulnerabilidad.
- Estas técnicas se combinan frecuentemente para lograr un mayor nivel de protección.
- Durante las fases de concepción y diseño de un tratamiento de datos de carácter personal, se debe determinar de forma precisa los márgenes adecuados de generalización y supresión, dentro de unos límites razonables que impidan la distorsión de la realidad y mantengan un determinado grado de precisión y fidelidad.

2.3. Conclusiones

Para un hospital de tamaño medio y a la vista del estudio realizado de las cinco soluciones orientadas al análisis de datos de farmacovigilancia, podemos concluir que ninguna de ellas se adapta a los requisitos necesarios para la realización de estudios sobre posibles RAM en pacientes polimedicados empleando un repositorio externo anonimizado. En este sentido, todas las soluciones adolecen de una serie de limitaciones comunes que paso a exponer a continuación:

- **Acceso a datos:** Están diseñadas para grandes bases de datos de farmacovigilancia pública, no para trabajar directamente con datos clínicos internos.
- **Nivel de análisis:** Están más orientadas a la detección de señales poblacionales (asociaciones medicamento-RAM globales) que a analizar casos individuales en profundidad.
- **Adaptabilidad:** Sería necesario desarrollar interfaces o adaptaciones específicas para importar y analizar datos del hospital.
- **Coste y complejidad:** PV-Works y, en menor medida, OpenVigil y VigiMine, pueden requerir recursos técnicos que un hospital de tamaño medio tal vez no tenga disponible fácilmente.
- **Privacidad:** Aunque el repositorio sea anonimizado, integrar datos sensibles en sistemas externos requiere controles de seguridad extra.

De todas las soluciones analizadas, solamente REDCap sería probablemente la mejor base para capturar y manejar datos retrospectivos de los historiales clínicos de los pacientes, pero REDCap no realiza un proceso ETL por sí mismo, por lo que sería necesario desarrollar adaptaciones externas específicas para importar datos en un repositorio externo anonimizado, así como complementarlo con herramientas de análisis estadísticos para detectar patrones de RAM en pacientes polimedicados. Por otra parte, las herramientas como OpenVigil, VigiMine y VigiRank serían útiles como apoyo para contrastar resultados, pero no como como plataformas principales de análisis interno.

A continuación, en la siguiente tabla se muestran las principales limitaciones de cada una de las herramientas:

Tabla 5. Limitaciones de las herramientas.

Herramienta	Limitaciones principales
OpenVigil	Diseñada para datos públicos FAERS, no para repositorios hospitalarios. Trabaja con datos agregados, no analiza historias clínicas individuales. Personalización limitada sin programación adicional.
PV-Works	Pensada para farmacovigilancia en la industria farmacéutica. Alto coste de implementación y mantenimiento en un hospital de tamaño medio.
VigiMine	Se basa en señales poblacionales de VigiBase, no permite análisis directos de datos hospitalarios. No trabaja con datos clínicos individuales del hospital.
VigiRank	Detecta señales RAM mediante ranking pero sobre grandes bases de datos (VigiBase). No está diseñada para trabajar con datos locales específicos ni para análisis de cohortes hospitalarias.
REDCap	Excelente para recolectar y gestionar datos clínicos, pero no incorpora nativamente algoritmos avanzados de farmacovigilancia. No incorpora un proceso ETL por sí mismo (requiere de adaptación externa específica). Requiere de módulos externos para análisis de patrones RAM. Necesita de integración personalizada con otros sistemas para mejorar la detección de señales.

Fuente: Elaboración propia, 2025.

3. Objetivos y metodología de trabajo

Este trabajo propone el desarrollo de una solución adaptada a las necesidades específicas de los profesionales de la UF de un hospital de tamaño medio para la realización de estudios retrospectivos relacionados con RAM en pacientes polimedicados de edad avanzada. Cada objetivo específico está diseñado con la finalizad de ser alcanzable, medible y comprobable para garantizar su contribución con el objetivo principal. La metodología del trabajo está organizada en seis fases esenciales, desde la conceptualización hasta la implementación final con el propósito de proporcionar una herramienta eficaz para la UF.

3.1. Objetivo general

El propósito general del trabajo consiste en diseñar e implementar una arquitectura de procesamiento de datos que permita la anonimización, almacenamiento y análisis de información clínica sintética, orientada a la detección de patrones asociados a reacciones adversas a medicamentos, mediante la aplicación de técnicas ETL, anonimización y consultas analíticas en un entorno relacional.

3.2. Objetivos específicos

1. Diseñar y simular un proceso ETL que genere un conjunto datos sintéticos, aplique un protocolo de anonimización y cargue los datos resultantes en un repositorio externo anonimizado.
2. Diseñar e implementar un protocolo de anonimización sobre datos sintéticos que preserve la utilidad analítica para estudios sobre RAM.
3. Desarrollar e implementar un repositorio externo anonimizado empleando el sistema de gestión de bases de datos PostgreSQL, preparado para recibir y almacenar los datos transformados por el proceso ETL.
4. Diseñar e implementar un conjunto de consultas analíticas básicas sobre el repositorio externo anonimizado, con el fin de extraer indicadores clave sobre la relación entre polimedicación y RAM en pacientes polimedicados (≥5 medicamentos) pertenecientes a grupos etarios de alto riesgo (≥65 años).

5. Explorar la generación de evidencia local mediante el análisis de combinaciones farmacológicas asociadas a ingresos hospitalarios potencialmente evitables, facilitando la detección de señales de riesgo clínico.

3.3. Metodología de trabajo

Este trabajo ofrece una solución tecnológica para realizar estudios retrospectivos, de carácter descriptivo, basado en datos sintéticos de pacientes polimedicados de edad avanzada atendidos por el servicio de urgencias hospitalarias.

Los estudios retrospectivos pretenden identificar la frecuencia y características de las RAM mediante el análisis de registros clínicos anonimizados, siguiendo una estrategia que garantice tanto la utilidad científica como el cumplimiento legal y ético, especialmente con el *Reglamento General de Protección de Datos (RGPD)* y la *Ley Orgánica 3/2018 de Protección de Datos y Garantía de los Derechos Digitales (LOPDGDD)*. La investigación deberá contar con la autorización del *Comité de Ética de Investigación Clínica (CEIC)* del hospital y con el aval del *Delegado de Protección de Datos (DPD)*. Además, se trasladará un acuerdo de uso tanto al Departamento de Sistemas de Información (DSI) como al equipo investigador de la UF.

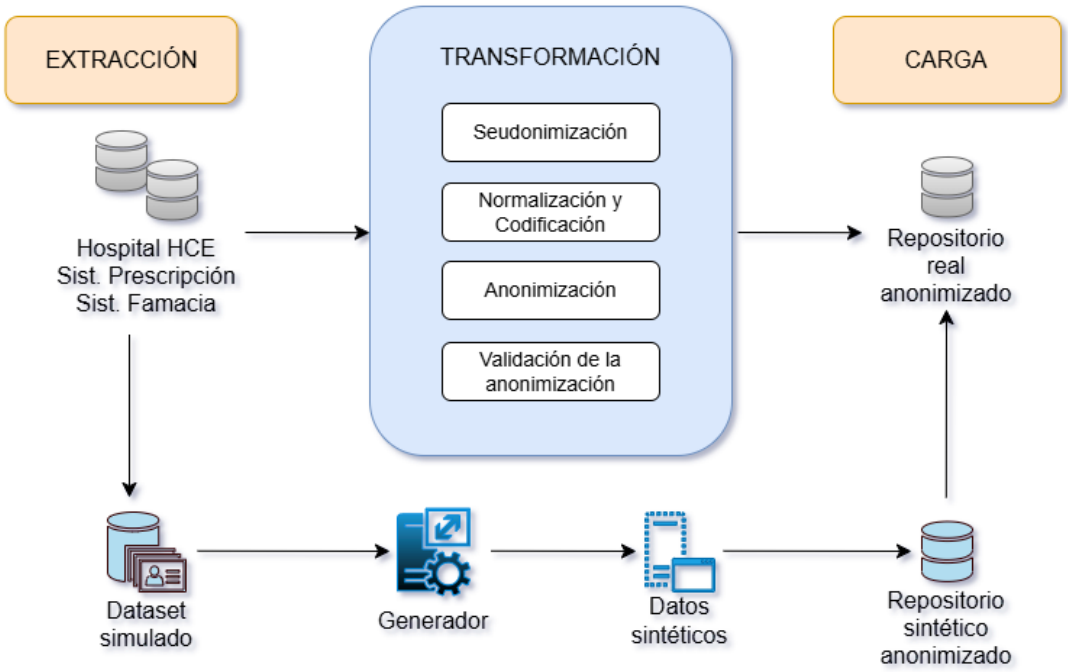
Aunque este proyecto podría haberse abordado en su conjunto, únicamente se centró en la simulación de un proceso ETL y en el desarrollo e implementación de un repositorio sintético anonimizado, abordándose mediante dos líneas de trabajo paralelas bien diferenciadas.

Como punto de partida y de forma conjunta, en una primera fase se abordaron con el DSI y la UF los objetivos específicos del repositorio, los sistemas fuente disponibles en el hospital y el dataset necesario para el subproceso de extracción del ETL. Además, se definieron los criterios de inclusión/exclusión de datos, la transformación de estos y cómo el ETL debía entregar los datos una vez preparados para la carga en el repositorio sintético anonimizado.

Para acelerar la ejecución del proyecto, se diseñó y se desarrolló paralelamente un repositorio externo con datos sintéticos anonimizados desde el principio, lo que permitió desacoplar su desarrollo de los datos reales del hospital, simular la salida del ETL empleando un generador de datos sintéticos con estructuras idénticas a las reales, validar las reglas de anonimización y realizar consultas analíticas básicas sobre el repositorio externo anonimizado.

El beneficio de este enfoque anticipado ofrece mayor flexibilidad para simular estructuras, validar modelos, probar flujos sin depender del acceso a datos reales, reducir riesgos al poder validar la utilidad de los datos y acelerar la puesta en producción del repositorio real anonimizado.

Figura 4. Proceso de creación de un repositorio externo anonimizado.



Fuente: Elaboración propia, 2025

3.3.1. Fase 1: Análisis de requisitos y necesidades de información

En primer lugar, se llevó a cabo un análisis de requisitos y necesidades del equipo investigador de la UF, se definieron unos objetivos claros, se identificaron las variables clave del estudio, las métricas de interés, y las especificaciones del sistema para una correcta integración de los datos generados por el ETL real en el repositorio real anonimizado.

3.3.2. Fase 2: Diseño conceptual del repositorio sintético anonimizado

En esta fase se creó una estructura de base de datos relacional orientada a consultas retrospectivas sobre RAM, modelando el diagrama Entidad-Relación para representar la parte estática como las tablas, las relaciones entre tablas, el establecimiento de claves, tipos de datos y campos sensibles a anonimizar.

3.3.3. Fase 3: Simulación de la fuente de datos

En esta etapa, ante la falta de datos reales, se emplearon datos sintéticos realistas y coherentes con la casuística clínica. Se definió el esquema de datos sintéticos como el tipo de variables, valores esperados, proporciones, etc. Para obtener un conjunto de datos sintéticos se empleó tecnología de generación de datos simulados que respetaran la lógica del estudio sin utilizar información real.

3.3.4. Fase 4: Diseño del proceso de anonimización

El objetivo de esta fase consistió en aplicar un protocolo de anonimización al conjunto de datos sintéticos como si fuesen reales, por lo que se identificaron qué atributos eran identificadores directos, cuáles eran cuasi-identificadores o identificadores indirectos y se aplicaron técnicas de anonimización para obtener una versión anonimizada del repositorio sintético listo para la realización de pruebas.

3.3.5. Fase 5: Implementación del repositorio de pruebas

En esta fase se construyó de forma local un repositorio sintético funcional para la realización de pruebas y validación, llevando a cabo actividades como la creación del repositorio en el gestor de bases de datos PostgreSQL, la carga del dataset anonimizado, así como consultas básicas simuladas como la búsqueda de RAM, el cruce de tratamientos y RAM, análisis por grupos de edad, etc. Con todo ello obtuvimos un repositorio funcional, ejecutable en un entorno local y listo para su evaluación.

3.3.6. Fase 6: Evaluación y validación

El objetivo de esta fase fue verificar que el repositorio sintético cumplía con los objetivos del proyecto, así como realizar una serie de actividades:

- Validar la estructura de la base de datos relacional verificando la integridad de las relaciones y su rendimiento.
- Validar la lógica para comprobar si las consultas arrojaban resultados clínicamente admisibles.
- Analizar la utilidad del repositorio y sus funcionalidades para verificar si respondía a las necesidades del equipo investigador.

4. Desarrollo e implementación

Este capítulo describe el proceso de desarrollo e implementación del repositorio externo con datos sintéticos orientado a la realización de estudios clínicos retrospectivos sobre RAM. Ante la imposibilidad de acceder inicialmente a datos reales por razones éticas y de privacidad, se optó por el uso de datos sintéticos, replicando la estructura y la lógica clínica de los datos hospitalarios previstos en el diseño del proceso ETL real.

A lo largo del capítulo se detallan las decisiones adoptadas en el diseño del modelo de datos, la construcción de las tablas y relaciones, la selección de tecnologías o herramientas para la generación sintética de datos, así como las estrategias de anonimización simulada y carga en el repositorio. Este entorno de pruebas permitió validar la utilidad del modelo para futuros estudios y asegurar la coherencia de las consultas analíticas previstas por el equipo investigador, desacoplando el desarrollo del repositorio del acceso a los datos reales y garantizando la continuidad de posibles mejoras futuras.

4.1. Análisis de requisitos y necesidades de información

4.1.1. Análisis de necesidades de la Unidad de Farmacología

El punto de partida fundamental para el diseño del repositorio externo, aunque sea en su versión sintética anonimizada, consistió en la identificación de las necesidades del equipo investigador. Teniendo en cuenta que el repositorio debía dar soporte a estudios retrospectivos sobre RAM en pacientes polimedicados (≥ 5 fármacos) de edad avanzada (≥ 65 años) atendidos por el servicio de urgencias, fue imprescindible captar y comprender qué tipo de información clínica se requería, cómo debía estar estructurada y qué funcionalidades de análisis se necesitaban.

Durante la fase de análisis se mantuvieron reuniones conjuntas con el DSI del hospital y con los profesionales clínicos, principalmente farmacólogos e investigadores de la UF con el objetivo de traducir sus preguntas en requisitos técnicos. Este proceso permitió recoger información clave sobre:

- **Variables clínicas de interés**: edad, género, medicamentos implicados en el momento del ingreso, códigos de RAM y gravedad registrados, así como el tiempo de hospitalización (fecha alta – fecha ingreso).
- **Criterios de selección de pacientes**: inclusión de pacientes de edad avanzada atendidos por el servicio de urgencias en un periodo definido y que estuviesen polimedicados (es decir, que en el momento del ingreso siguieran un tratamiento con cinco o más principios activos de forma simultánea).
- **Relaciones causales**: la necesidad de detectar combinaciones de tratamientos farmacológicos con eventos adversos.
- **Necesidades de agregación y análisis estadístico**: necesidad de poder calcular tasas de RAM, frecuencia de medicamentos implicados, patrones de RAM y segmentación por rango de edad y/o género.
- **Privacidad y ética**: aunque los datos utilizados fuesen sintéticos, se definieron desde el principio los atributos sensibles que deberían ser objeto de anonimización para simular fielmente el proceso que se fuese a llevar a cabo en el entorno real.

En esta fase de análisis de necesidades, se confeccionó un documento de requisitos funcionales que sirvió de base para el diseño conceptual del repositorio y la generación de los datos sintéticos. Asimismo, permitió establecer un lenguaje común entre clínicos y técnicos, facilitando una colaboración eficaz bajo una lógica de desarrollo iterativo.

Un aspecto clave de esta fase fue la validación de las consultas clínicas básicas como, por ejemplo:

- "¿Qué combinaciones de fármacos están asociadas con mayor riesgo de RAM en pacientes polimedicados (≥ 5 medicamentos) y edad avanzada (≥ 65 años) teniendo en cuenta el grado de solapamiento temporal de los tratamientos?".
- "¿Cuál es la frecuencia de RAM sobre ingresos hospitalarios en pacientes polimedicados frente a no polimedicados?".
- "¿Cuál es la frecuencia de RAM sobre ingresos hospitalarios en pacientes polimedicados (≥ 5 medicamentos) según rango de edad (entre 65-74 años; entre 75-84 años; mayores ≥ 85 años)?".

Estas consultas se consideraron representativas y sirvieron para orientar la arquitectura del repositorio, el diseño del esquema de datos y las pruebas funcionales con datos sintéticos.

En definitiva, este subapartado del proyecto evidencia que el diseño técnico del sistema partió de una base sólida y realista, basada en las necesidades concretas del equipo investigador, y permitió asegurar que el entorno desarrollado fuese funcional y útil para el tipo de estudios clínicos que se llevaron a cabo posteriormente.

4.1.2. Finalidad del entorno sintético

La creación de un entorno sintético tiene como objetivo principal permitir el desarrollo anticipado y autónomo del repositorio externo de investigación, sin depender del acceso a los datos clínicos reales del hospital. En el contexto de este trabajo, donde se pretende construir una base de datos anonimizada que permita realizar estudios retrospectivos sobre RAM en pacientes polimedicados de edad avanzada y atendidos por el servicio de urgencias hospitalarias, el uso de datos reales está condicionado a una serie de autorizaciones éticas, legales y organizativas que requieren tiempo y coordinación con distintos departamentos del hospital. Ante esta limitación inicial, se valoró y se optó por generar un entorno de pruebas basado en datos sintéticos estructurados que simularan con realismo la complejidad y organización de los datos clínicos reales.

Este entorno sintético permitió avanzar en fases clave del desarrollo del sistema, como el diseño lógico del repositorio, la validación de las estructuras de datos, la ejecución de consultas clínicas simuladas, la implementación de reglas de anonimización, así como la prueba de mecanismos de exportación y explotación de la información. Además, facilitó la colaboración con el equipo investigador, permitiéndoles validar las variables, relaciones y estructuras necesarias para responder a sus preguntas desde el inicio del proyecto, sin exponer ningún dato sensible.

Por otro lado, el entorno sintético sirvió como marco de validación del modelo conceptual y funcional que fue posteriormente utilizado en el entorno real, asegurando que el diseño adoptado fuera coherente, viable y adaptable al flujo ETL real desarrollado en paralelo por el DSI. En este sentido, se convirtió en una herramienta que desacoplaba los tiempos de

desarrollo técnico del repositorio de los tiempos administrativos y técnicos asociados al acceso a los datos reales, acelerando de este modo el desarrollo del proyecto.

En resumen, el entorno sintético cumplió una doble función, por un lado, es un entorno funcional que permitió construir y validar el repositorio externo en condiciones seguras; por otro, es un entorno pedagógico y exploratorio que contribuye al aprendizaje, la iteración ágil del diseño, y la preparación para la integración futura con datos reales anonimizados.

4.1.3. Alcance funcional del repositorio

El repositorio externo con datos sintéticos se diseñó como una plataforma de almacenamiento y consulta estructurada, orientada a facilitar estudios retrospectivos simulados sobre RAM. Aunque este tipo de entornos no contiene información real, su alcance funcional estaba alineado con las necesidades investigadoras y los requerimientos del sistema final que se implementaron con datos anonimizados provenientes historias clínicas del hospital.

El repositorio cumple varias funciones clave:

En primer lugar, ofrece una estructura relacional robusta, en la que se reflejan las relaciones clínicas entre entidades y es coherente con el modelo de datos previsto en un entorno real, permitiendo su puesta en producción en un corto espacio de tiempo.

En segundo lugar, el repositorio permite la ejecución de consultas analíticas complejas, diseñadas para dar respuesta a consultas sobre la incidencia de RAM en función del número de medicamentos implicados en los tratamientos, la aparición de eventos adversos en distintos grupos de edad o la distribución temporal de las visitas a urgencias. Estas consultas se prueban sobre los datos sintéticos anonimizados para validar su estructura, tiempos de ejecución y capacidad de extracción de conocimiento.

Asimismo, en el entorno sintético se incluyen mecanismos que simulan de forma realista los procesos de anonimización, de manera que, aunque los datos sintéticos no contienen información real, el proceso de anonimización se implementa como un ejercicio previo para garantizar que las técnicas y herramientas seleccionadas serán efectivas cuando se apliquen al entorno real.

Otro aspecto relevante dentro del alcance funcional es la capacidad del repositorio para integrarse con herramientas de análisis estadístico o sistemas externos de visualización y explotación de datos. Por ello, resulta interesarte asegurar la portabilidad de la información mediante formatos interoperables (por ejemplo, CSV, JSON o SQL) con la finalidad de preparar el sistema para su conexión futura con entornos de análisis clínico.

Por último, este repositorio tiene una función transversal como entorno de validación y aprendizaje. Permite que tanto los desarrolladores como los investigadores puedan familiarizarse con el modelo de datos, las reglas de explotación y los mecanismos de protección de la privacidad sin poner en riesgo la privacidad de ningún paciente real. Además, funciona como plataforma de pruebas para los procesos de carga, verificación de integridad y control de calidad del dato que se implementarán posteriormente en el entorno real.

4.1.4. Justificación del uso de datos sintéticos

El uso de datos sintéticos en este trabajo de fin de grado responde a una necesidad práctica, ética y metodológica para avanzar en el diseño e implementación de un repositorio externo de datos clínicos sin depender de la disponibilidad inmediata de datos reales, cuya manipulación está sujeta a estrictas restricciones legales, normativas y de confidencialidad.

Ante esta realidad, el uso de datos sintéticos surge como una solución para desacoplar el desarrollo técnico del repositorio del acceso efectivo a los datos reales. Los datos sintéticos son generados artificialmente, ya sea de forma aleatoria o a partir de patrones estadísticos derivados de estructuras reales (cuando estas están disponibles), y permiten simular de manera realista la organización, relaciones y distribución de los datos clínicos sin comprometer la privacidad de ningún paciente.

Desde el punto de vista funcional, los datos sintéticos permiten validar con antelación el diseño del modelo de datos, verificar las relaciones entre entidades clínicas, implementar y ensayar mecanismos de anonimización, probar consultas SQL e incluso realizar pruebas de exportación e integración en herramientas de análisis. Su uso reduce drásticamente los tiempos de desarrollo y elimina el riesgo legal facilitando una metodología ágil de desarrollo.

Además, los datos sintéticos ofrecen una ventaja pedagógica importante. Permiten que los equipos técnicos y clínicos trabajen de manera colaborativa sobre una base segura, sin miedo

a comprometer la confidencialidad. Esto promueve el aprendizaje compartido, la validación temprana de hipótesis y la construcción conjunta del entorno de investigación.

Cabe señalar que, aunque los datos sintéticos son útiles para validar aspectos estructurales y funcionales, adolecen de ciertas limitaciones debido a que pueden no reflejar con exactitud la complejidad clínica real, no contienen errores ni incoherencias propias de entornos sanitarios reales, y no pueden utilizarse para extraer conclusiones científicas sólidas. Por ello, su uso debe entenderse como complementario y temporal, destinado a preparar el terreno para una futura implementación con datos reales.

En resumen, la incorporación de datos sintéticos permite avanzar en paralelo al desarrollo del proceso ETL real del hospital, garantizar un entorno de pruebas seguro y eficaz, y facilitar la experimentación y validación de los elementos clave del sistema sin comprometer la privacidad ni ralentizar la ejecución del proyecto.

Figura 5. Ventajas de los datos sintéticos vs datos reales.

Fuente: datos.gov.es, 2023

4.2. Diseño conceptual del repositorio sintético anonimizado

El Diseño Conceptual es una representación de las propiedades estáticas que se necesitan para satisfacer los requerimientos del repositorio sintético anonimizado, en este caso los requerimientos son los analizados en el punto anterior. Es la primera descripción formal del repositorio, y es independiente del Sistema de Gestión de Bases de Datos (SGBD) que se vaya a utilizar.

El esquema conceptual que se elabora en este diseño incluye:

- un diagrama Entidad-Relación para representar la parte estática (objetos, atributos y dominios asociados a los atributos).
- las restricciones de integridad del repositorio que no se puedan expresar en el diagrama E-R y
- las transacciones que representan la parte dinámica de dicho repositorio.

4.2.1. Diagrama Entidad-Relación

A partir del análisis de la realidad y de los requerimientos de la UF, para la obtención del diagrama Entidad-Relación se van a seguir los siguientes pasos:

- Primero: Identificación de Entidades y Atributos.
- Segundo: Identificación de Relaciones entre Entidades.
- Tercero: Transacciones o Procesos sobre el Repositorio.

Identificación **de Entidades y Atributos**

Del análisis de necesidades de la UF se deducen las siguientes entidades:

- Los pacientes: "**Pacientes**"
- Los diagnósticos de RAM: "**Ram**"
- Los medicamentos: "**Medicamentos**"
- Los ingresos: "**Ingresos**"
- Las RAM de ingreso: **"Ram-Ingreso"**
- Los tratamientos: "**Tratamientos**"

A continuación, se especifican los atributos de cada objeto o entidad junto con su dominio asociado y una descripción de su significado:

Pacientes		
Atributo	**Dominio**	**Significado**
id_paciente	dom_ipac	"Nº secuencial"
genero	dom_genero	"Género del paciente ('M', 'F', 'O')"
rango_edad	dom_redad	"Grupo etario (1=65-74, 2=75-84, 3=≥85)"

Ram		
Atributo	**Dominio**	**Significado**
id_ram	dom_iram	"Nº secuencial"
codigo_cie	dom_cie	"Código de la reacción adversa"
reaccion_adversa	dom_rad	"Breve descripción de la reacción adversa"

Medicamentos		
Atributo	**Dominio**	**Significado**
id_medicamento	dom_imed	"Nº secuencial"
codigo_atc	dom_atc	"Código del principio activo (ej. N05BA01)"
principio_activo	dom_prinac	"Nombre del principio activo"

Ingresos		
Atributo	**Dominio**	**Significado**
id_ingreso	dom_ingr	“Nº secuencial”
fecha_ingreso	dom_fch	“Fecha de ingreso”
fecha_alta	dom_fch	“Fecha de alta”

Ram_Ingreso		
Atributo	**Dominio**	**Significado**
id_ingreso	dom_ingr	“Número identificativo del ingreso”
id_ram	dom_iram	“Número identificativo de la ram”
gravedad	dom_grav	“Gravedad de la ram ('Leve', 'Moderada', 'Grave')”

Tratamientos		
Atributo	**Dominio**	**Significado**
id_ingreso	dom_ingr	“Número identificativo del ingreso”
id_medicamento	dom_med	“Número identificativo del medicamento”
fecha_inicio	dom_fch	“Inicio del tratamiento del medicamento”
fecha_fin	dom_fch	“Fin del tratamiento del medicamento”
dosis	dom_dosis	“Dosis de principio activo (‘Baja’, ‘Media’, ‘Alta’)”

A continuación, se especifica el tipo de datos asociado a cada nombre de dominio:

Nombre-Dominio	Tipo de Datos
dom_atc	varchar(7)
dom_cie	varchar(10)
dom_dosis	(‘Baja’, ‘Media’, ‘Alta’)
dom_fch	date
dom_genero	(‘M’, ‘F’, ‘O’)
dom_grav	('Leve', 'Moderada', 'Grave')
dom_imed	entero positivo
dom_ingr	entero positivo
dom_ipac	entero positivo
dom_iram	entero positivo
dom_prinac	text
dom_rad	text
dom_redad	[1..3]

Identificación de Relaciones entre Entidades

Una relación queda definida con:

- el nombre de las entidades que se relacionan,
- las cardinalidades máximas y mínimas de la participación de cada entidad sobre las demás en la relación,
- y los atributos propios de la relación.

Relación entre Pacientes e Ingresos: **"Tiene"**

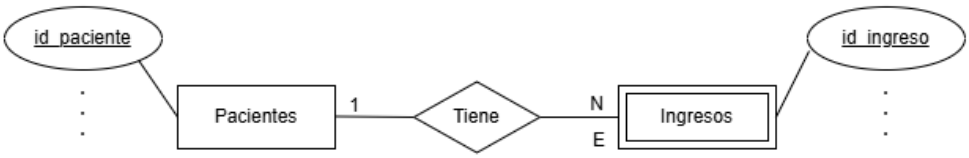
Un paciente puede tener varios ingresos o ninguno.

- Card. mín. de Pacientes sobre Ingresos: 0 (sin restricción de existencia).
- Card. Máx. de Pacientes sobre Ingresos: N.

Todo ingreso se relaciona con un solo paciente.

- Card. mín. de Ingresos sobre Pacientes: 1 (restricción de existencia).
- Card. Máx. de Ingresos sobre Pacientes: 1

No se consideran los atributos propios de la relación.



Relación entre Ram e Ingresos: **"Ram-Ingreso"**

Una ram puede aparecer en varios ingresos o ninguno.

- Card. mín. de Ram sobre Ingresos: 0 (sin restricción de existencia).
- Card. Máx. de Ram sobre Ingresos: N.

En un ingreso pueden aparecer varias ram o al menos una.

- Card. mín. de Ingresos sobre Ram: 1 (restricción de existencia).
- Card. Máx. de Ingresos sobre Ram: N

No se consideran los atributos propios de la relación.

Relación entre Ingresos y Tratamientos: **"Contiene"**

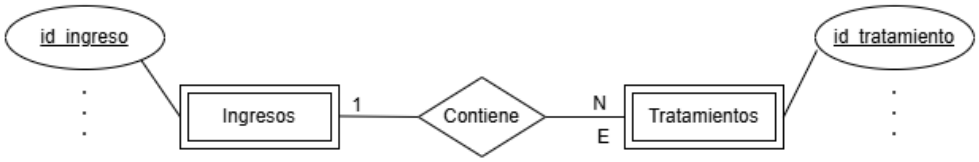
Un ingreso puede contener varios tratamientos o ninguno.

- Card. mín. de Ingresos sobre Tratamientos: 0 (sin restricción de existencia).
- Card. Máx. de Ingresos sobre Tratamientos: N.

Todo tratamiento se relaciona con un solo ingreso.

- Card. mín. de Tratamientos sobre Ingresos: 1 (restricción de existencia).
- Card. Máx. de Tratamientos sobre Ingresos: 1

No se consideran los atributos propios de la relación.



Relación entre Medicamentos y Tratamientos: **"Med-Trat"**

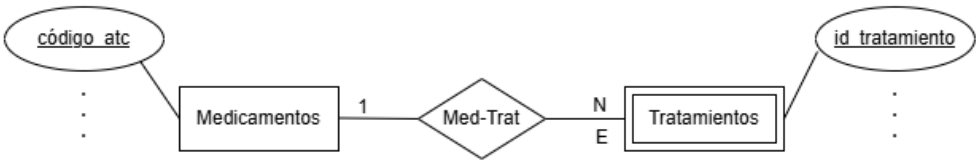
Un medicamento puede aparecer en varios tratamientos o ninguno.

- Card. mín. de Medicamentos sobre Tratamientos: 0 (sin restricción de existencia).
- Card. Máx. de Medicamentos sobre Tratamientos: M

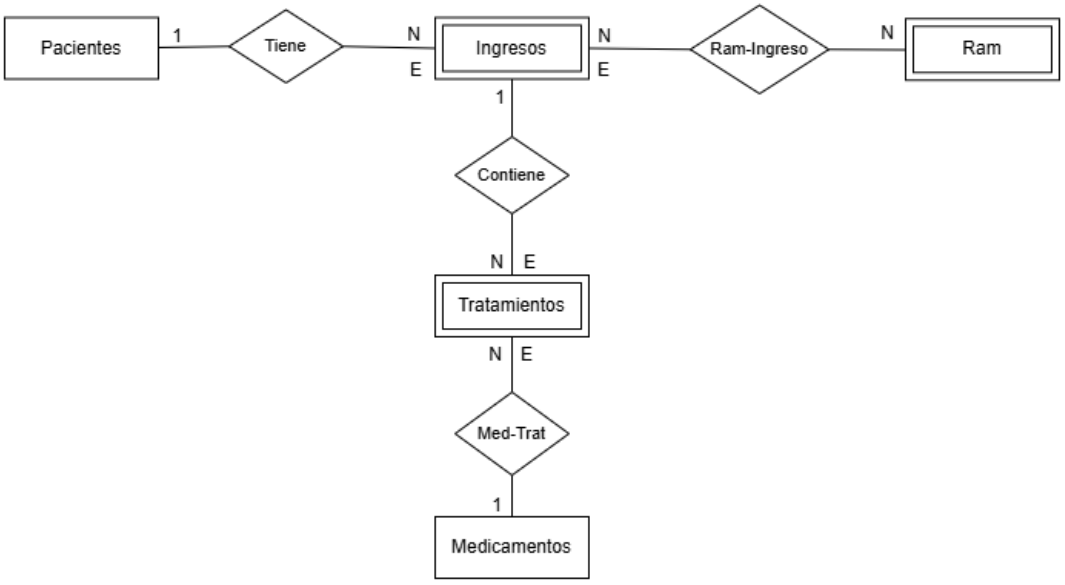
Todo tratamiento se relaciona con un solo medicamento.

- Card. mín. de Tratamientos sobre Medicamentos: 1 (restricción de existencia).
- Card. Máx. de Tratamientos sobre Medicamentos: 1

No se consideran los atributos propios de la relación.



A continuación, se presenta el Diagrama Entidad-Relación sin especificar atributos.



Transacciones o Procesos sobre el Repositorio

Para cada objeto (entidad o relación) del diagrama Entidad-Relación se diseñan las transacciones básicas que especifican los procesos que se llevan a cabo sobre dichos objetos: inserción, borrado, modificación y consulta de sus ocurrencias. En este trabajo, como el repositorio permanecerá invariable una vez realizada la carga de datos mediante el proceso ETL, sólo nos centraremos en las transacciones de inserción y en otras transacciones que representan requerimientos dinámicos de acuerdo con las necesidades del repositorio y del usuario.

Las transacciones de inserción se han especificado con lenguaje natural para facilitar la comprensión, manteniendo una estructura bien definida: "*Datos necesarios para la transacción*", "*Condiciones*" y "*Acciones*", si se cumplen todas las condiciones, se realizarán todas las acciones de la transacción. Por otra parte, se han estudiado 'otras transacciones' como las consultas y procesos no básicos más utilizados sobre el repositorio. Por lo último, antes del diseño de cada transacción se estudian las restricciones a las que están sometidas las entidades o relaciones afectadas.

- **Transacciones de Inserción:**

La inserción en entidades siempre está condicionada por la restricción de integridad representada por el atributo identificador: "*no puede contener un valor nulo y no debe existir otra ocurrencia con el mismo valor en el atributo identificador*". Esta restricción siempre se expresará en las condiciones de la transacción, junto con las posibles restricciones de valor no nulo definidas sobre atributos descriptores de la entidad.

Pacientes: No tiene restricciones. Para insertar o dar de alta un nuevo paciente no hace falta ningún requisito especial, únicamente que se disponga de un id_paciente que no se corresponda a otro paciente.

TRANSACCIÓN insertar_ paciente

DATOS NECESARIOS:

‘ los atributos de la nueva ocurrencia de paciente,

CONDICIONES:

‘ que el valor del atributo identificador no sea nulo,

‘ que no exista otro paciente con el mismo valor en el atributo identificador o clave,

ACCIONES:

‘ insertar en Pacientes la nueva ocurrencia.

Medicamentos: No tiene restricciones. Para insertar o dar de alta un nuevo medicamento no hace falta ningún requisito especial, únicamente que se disponga de un codigo_atc que no se corresponda a otro medicamento.

TRANSACCIÓN insertar_ medicamento

DATOS NECESARIOS:

‘ los atributos de la nueva ocurrencia de medicamento,

CONDICIONES:

‘ que el valor del atributo identificador no sea nulo,

‘ que no exista otro medicamento con el mismo valor en el atributo identificador o clave,

ACCIONES:

‘ insertar en Medicamentos la nueva ocurrencia.

Ram: No tiene restricciones. Para insertar o dar de alta una nueva ram no hace falta ningún requisito especial, únicamente que se disponga de un codigo_cie10 que no se corresponda a otra ram.

TRANSACCIÓN insertar_ ram

DATOS NECESARIOS:

‘ los atributos de la nueva ocurrencia de ram,

CONDICIONES:

‘ que el valor del atributo identificador no sea nulo,

‘ que no exista otra ram con el mismo valor en el atributo identificador o clave,

ACCIONES:

‘ insertar en Ram la nueva ocurrencia.

Ingresos: tiene restricciones de existencia sobre las relaciones Tiene y Ram-Ingreso. Cuando se inserte un nuevo ingreso se deberá asociar con un paciente y la ram con la que ingresa.

TRANSACCIÓN insertar_ingreso

DATOS NECESARIOS:

‘ los atributos de la nueva ocurrencia de ingreso,

‘ el atributo identificador de la ocurrencia de paciente con el que se relacione en Tiene,

CONDICIONES:

‘ que el valor del atributo identificador de la nueva ocurrencia de ingreso no sea nulo y no aparezca en otra ocurrencia de Ingresos,

‘ que exista la ocurrencia de paciente con la que se va a relacionar en Tiene,

ACCIONES:

‘ insertar en Ingresos la nueva ocurrencia,

‘ insertar en Tiene una ocurrencia que relacione la nueva ocurrencia de Ingresos con el paciente indicado,

Tratamientos: tiene restricciones de existencia sobre las relaciones Contiene y Med-Trat. Cuando se inserte un nuevo tratamiento se deberá asociar con un ingreso y con un medicamento.

TRANSACCIÓN insertar_tratamiento

DATOS NECESARIOS:

‘ los atributos de la nueva ocurrencia de tratamiento,

‘ los atributos identificadores de las ocurrencias de ingreso y medicamento con las que se relacione en Contiene y Med-Trat,

CONDICIONES:

‘ que el valor de los atributos identificadores de la nueva ocurrencia de tratamiento no sean nulos y no aparezcan en otra ocurrencia de Tratamiento,

‘ que existan las ocurrencias de ingreso y medicamento con la que se va a relacionar en Contiene y Med-Trat,

ACCIONES:

‘ insertar en Tratamientos la nueva ocurrencia,

‘ insertar en Contiene una ocurrencia que relacione la nueva ocurrencia de Tratamientos con el ingreso indicado,

‘ insertar en Med-Trat una ocurrencia que relacione la nueva ocurrencia de Tratamientos con el medicamento indicado.

TRANSACCIÓN insertar_ram_ingreso

DATOS NECESARIOS:

‘ los atributos de la nueva ocurrencia de ram_ingreso,

‘ los atributos identificadores de las ocurrencias de ingreso y ram con las que se relacione en Ram-Ingreso,

CONDICIONES:

‘ que el valor de los atributos identificadores de la nueva ocurrencia de Ram-Ingreso no sean nulos y no aparezcan en otra ocurrencia de Ram-Ingreso,

‘ que existan las ocurrencias de ingreso y ram con la que se va a relacionar en Ram-Ingreso,

ACCIONES:

‘ insertar en Ram-Ingreso la nueva ocurrencia,

‘ insertar en Ram-Ingreso una ocurrencia que relacione la nueva ocurrencia de Ram-Ingreso con el ingreso indicado,

‘ insertar en Ram-Ingreso una ocurrencia que relacione la nueva ocurrencia de Ram-Ingreso con la ram indicada.

- **Otras transacciones:**

Además de las transacciones básicas de inserción, también son necesarias otras transacciones de consulta para completar los requerimientos dinámicos o los procesos que se llevan a cabo sobre los objetos del diagrama Entidad-Relación.

Así pues, del análisis de necesidades de la UF se deducen los siguientes procesos además de los ya vistos:

- Consulta 1: “¿Qué combinaciones de fármacos están asociadas con mayor riesgo de RAM en pacientes polimedicados (≥ 5 fármacos) y edad avanzada (≥ 65 años) teniendo en cuenta el grado de solapamiento temporal de los tratamientos?”.
- Consulta 2: “¿Cuál es la frecuencia de RAM sobre ingresos hospitalarios en pacientes polimedicados frente a no polimedicados?”.
- Consulta 3: “¿Cuál es la frecuencia de RAM sobre ingresos hospitalarios en pacientes polimedicados (≥ 5 fármacos) según rango de edad (entre 65-74 años; entre 75-84 años; edad ≥85 años)?”.
- Consulta 4: “¿Cuál es la prevalencia de RAM en función del número de medicamentos y gravedad de la RAM?”.
- Consulta 5: “¿Cuál es la prevalencia de RAM por género?”.

4.2.2. Modelo Lógico de Datos

El modelo lógico de datos constituye la representación de las propiedades estructurales (estáticas) de la información necesarias para satisfacer los requerimientos del repositorio, *descrita en términos de un modelo de SGBD*, permitiendo organizar la información clínica de forma coherente, normalizada y orientada al análisis de RAM en pacientes polimedicados de edad avanzada.

El modelo de datos escogido en este trabajo es el *modelo relacional*. Así, el Esquema Lógico desarrollado en este punto ha consistido en la traducción del Esquema Conceptual a las estructuras del modelo relacional:

- Por un lado, los objetos del diagrama Entidad-Relación (parte estática del repositorio), en relaciones,
- y por otro lado, las transacciones del Esquema Conceptual, a transacciones del Esquema Lógico.

Transformación del Diagrama E-R a Relaciones:

La potencia del SGBD se muestra entre otras cosas por la capacidad de definir claves primarias (CP), claves alternativas (CAlt) y claves ajenas (CAj) en las relaciones, y restricciones de valor no nulo (VNN) en atributos. En este punto se van a utilizar estas definiciones y aunque muchas de estas nos las ofrecerá el SGBD elegido, otras deberán ser controladas por el programador.

La transformación de las entidades y relaciones del Diagrama Entidad-Relación en esquemas relacionales, se ha realizado siguiendo unas reglas determinadas de las que aquí sólo se van a exponer las utilizadas en este trabajo:

• Cada entidad del E-R se traduce en una relación con tantos atributos como tenía la entidad. El dominio asociado a los atributos es el mismo que se especificó en el Esquema Conceptual. La clave primaria de la relación la forman los atributos identificadores de la entidad, es decir; los que se subrayaron al nombrar los atributos de cada entidad. Los atributos con restricción de valor no nulo se definen especificando dicha restricción.

• Las relaciones binarias de cardinalidad 1:M no forman una relación aparte, sino que se añaden a la relación obtenida de la entidad de cardinalidad M el atributo identificador de la relación de la otra entidad (junto con los posibles atributos propios de la relación) definiéndose como clave ajena que referencia a aquella. Si además la entidad de cardinalidad M tiene una restricción de existencia sobre la relación, se define sobre el atributo identificador añadiéndole una restricción de valor no nulo.

Teniendo en cuenta las reglas anteriores, las relaciones que resultan de la traducción del Esquema Conceptual son las siguientes:

Pacientes (id_paciente: dom_ipac, genero: dom_genero, rango_edad: dom_redad)
CP: { id_paciente }
VNN: { rango_edad }
VNN: { genero }

Ingresos (id_ingreso: dom_ingr, id_paciente: dom_ipac, fecha_ingreso: dom_fch, fecha_alta: dom_fch, gravedad: dom_grav)
CP: { id_ingreso }
CAj: { id_paciente } referencia a Pacientes
VNN: { id_paciente }
VNN: { fecha_ingreso }
VNN: { fecha_alta }

Ram (id_ram: dom_iram codigo_cie: dom_cie, reacción_adversa: dom_rad)
CP: { id_ram }
VNN: { codigo_cie }

Medicamentos (id_medicamento: dom_imed, código_atc: dom_atc, principio_activo: dom_prinac)
CP: { id_medicamento }
VNN: { código_atc }

Tratamientos (id_ingreso: dom_ingr, id_medicamento: dom_imed, fecha_inicio: dom_fch, fecha_fin: dom_fch, dosis: dom_dosis)
CP: { id_ingreso } **CP**: { id_medicamento } **CAj:** { id_ingreso } referencia a Ingresos **CAj:** { id_medicamento } referencia a Medicamentos **VNN:** { id_ingreso } **VNN:** { id_medicamento } **VNN:** { fecha_inicio } **VNN:** { fecha_fin } **VNN:** { dosis }
Tratamiento activo en el ingreso: fecha_inicio <= fecha_ingreso

Ram_Ingreso (id_ingreso: dom_ingr id_ram: dom_iram gravedad: dom_grav)
CP: { id_ingreso } **CP**: { id_ram } **CAj:** { id_ingreso } referencia a Ingresos **CAj:** { id_ram } referencia a Ram **VNN:** { gravedad }

El diseño respeta las reglas de normalización para evitar redundancia y asegurar la integridad referencial. Cada objeto dispone de una clave primaria y de las correspondientes claves ajenas para mantener la trazabilidad entre entidades., además, todas las relaciones obtenidas ya se encuentran en 3ª Forma Normal, por lo que no es necesario modificarlas.

Transformación de las transacciones:

La traducción de las transacciones del Esquema Conceptual sobre el Esquema Relacional está en función de las herramientas que aporte el SGBD que se vaya a emplear. Es decir, si el SGBD controla la integridad de las claves primarias, entonces no será necesario comprobar en las "CONDICIONES" de las transacciones la restricción de integridad que introducen los atributos identificadores (que deberán ser únicos y no nulos). De la misma forma, si el SGBD proporciona la definición de claves ajenas y la restricción de valor no nulo en atributos, entonces tampoco es necesario comprobar en las transacciones del Esquema Relacional la integridad referencial que introducen las relaciones ni que los atributos no sean nulos si se han restringido así.

El SGBD elegido, el cual describiremos más adelante en el apartado 4.5.1 (Entorno y Tecnologías utilizadas), soporta las definiciones descritas anteriormente, por lo que el programador no tendrá que controlar por programa las definiciones anteriores.

Así pues, pasaremos a explicar las divergencias que surgen del cambio de las estructuras que utilizan los esquemas para representar la realidad:

- en el Esquema Conceptual (E.C.) se disponían de entidades, relaciones y atributos, mientras que en el Esquema Lógico (E.L.) la única estructura es la relación;
- en las entidades y relaciones se hablaba de ocurrencias, mientras que en las relaciones del E.L. se habla de tuplas, que no son otra cosa que una colección de pares (atributo, valor);
- las relaciones del E.L. agrupan, en ocasiones, a varios objetos del E.C., en concreto las relaciones binarias 1:M, las cuales se incluyen en la relación que representa a la entidad de cardinalidad M, por lo que las acciones que antes se hacían sobre varios objetos, ahora sólo se harán sobre uno.

4.2.3. Esquema Lógico de la Base de Datos

Representa la estructura organizativa y las relaciones lógicas de los datos dentro de la base de datos, independientemente de la implementación física y describe las tablas, campos, tipos

de datos, relaciones entre tablas y restricciones de integridad, definiendo cómo se organizan y relacionan los datos, sin importar dónde se almacenan. (*Ver Anexo B*).

4.2.4. Variables clínicas clave para el estudio

En este apartado se describen las variables clínicas esenciales para la realización de estudios retrospectivos de RAM en pacientes polimedicados de edad avanzada.

1.- Datos del Paciente (anónimos pero relevantes):

Variable	Descripción
rango_edad	Grupo etario: 1 (65–74), 2 (74–85), 3 (≥85). Crucial para estratificación.
genero	M, F, O. Puede haber diferencias en susceptibilidad a RAM.

2.- Datos del Ingreso hospitalario:

Variable	Descripción
fecha_ingreso	Permite establecer la cronología del evento RAM y del tratamiento.
fecha_alta	Define duración del ingreso y posibles complicaciones.
codigo_cie	Código de la RAM. Identifica la patología vinculada.
gravedad	Clasifica el impacto clínico (Leve, Moderada, Grave).

3.- Datos del Tratamiento farmacológico:

Variable	Descripción
codigo_atc	Código del medicamento según la clasificación ATC.
fecha_inicio	Inicio del tratamiento. Permite inferir coincidencia con la RAM.
fecha_fin	Fin del tratamiento. Evaluar duración e implicación temporal.
dosis	Dosis. Permite inferir coincidencia con la RAM.
Nº de medicamentos	Cálculo derivado para determinar si el paciente es polimedicado (≥5).

4.- Variables Derivadas / Analíticas (para estudios):

Variable	Descripción
es_polimedicado	TRUE/FALSE según si tiene ≥5 medicamentos activos al ingreso.
tiempo_hasta_ram	Diferencia entre fecha_inicio tratamiento y aparición de RAM.
frecuencia_ram_por_atc	Cuántas veces aparece una RAM tras uso de un medicamento específico.
combinacion_medicamentos	Conjunto de medicamentos activos en pacientes con RAM.
ram_por_rango_edad	Tasa de RAM segmentada por grupo etario.
ram_por_gravedad	Distribución de RAM según gravedad (para análisis de impacto clínico).

Este esquema lógico permite implementar consultas analíticas complejas, además, su estructura facilita la aplicación de reglas de anonimización, al mantener una separación clara entre identificadores clínicos y datos sensibles. En la versión sintética del repositorio, este modelo ha sido implementado con las mismas estructuras que se esperan en el entorno real, de modo que las pruebas, validaciones y consultas realizadas sean extrapolables y directamente reutilizables una vez se disponga de datos reales anonimizados.

4.3. Simulación de la fuente de datos

4.3.1. Criterios clínicos simulados

En la generación del dataset sintético estructurado se simularon distintos criterios clínicos con el objetivo de representar perfiles relevantes de pacientes, especialmente en situaciones de atención urgente y complejidad clínica. Los criterios seleccionados están alineados con condiciones frecuentemente observadas en pacientes con riesgo clínico elevado. A continuación, se detallan los criterios clínicos simulados:

1. **Pacientes**

- Edad: todos los pacientes deben pertenecer a un grupo etario ≥ 65 años.
- Género: la distribución debe ser realista, entre 47% - 49% Masculino ('M'), entre 49%-51% Femenino ('F'), entre 0,5% -2% Otro ('O')

2. **Ingresos**

- Distribución de ingresos: los pacientes pueden tener múltiples ingresos (simular 1–4 ingresos por paciente, con una distribución sesgada hacia 1–2 ingresos).
- Fecha de ingreso / alta: la fecha de ingreso debe ser menor que la fecha de alta (fecha_ingreso < fecha_alta), con una duración de hospitalización entre 2 y 15 días.
- Código RAM asociado: se aplica un código CIE que represente a una RAM.
- Gravedad: puede depender del tipo de RAM.
    - Leve: náuseas, cefalea, fatiga
    - Moderada: temblor muscular, hipotensión, disnea
    - Grave: anafilaxia, insuficiencia renal, hemorragia interna
    - Distribución de gravedad por edad/género: las personas entre 75-84 años pueden tener mayor riesgo de reacciones graves.

3. **Medicamentos**

- Código ATC: generar códigos válidos (e.g. A02BC01 para omeprazol, J01CA04 para amoxicilina).
- Nombre del principio activo: relacionarlo correctamente con el código ATC.

4. **Tratamientos**

- fecha_inicio ≤ fecha_ingreso (algunos tratamientos pueden haberse iniciado antes del ingreso y aparecer tratamientos temporalmente solapados en el momento del ingreso).
- fecha_fin ≤ fecha_alta (algunos tratamientos pueden terminar antes del alta, simular al azar con un 30% con fin anticipado).
- Número de tratamientos por ingreso: de 5 a 8, con una distribución realista en polimedicados ≥ 5 medicamentos. (media ≈ 5).
- No se repiten tratamientos de medicamentos dentro de un mismo ingreso de un paciente (Sin duplicados: id_ingreso, codigo_atc).
- Relación con RAMs:
    - Si hay una RAM de tipo cutáneo → antibióticos (J01), AINEs (M01)
    - Si hay una RAM hepática → estatinas (C10AA), antituberculosos (J04)
    - Si hay RAM hematológica → anticoagulantes y antitrombóticos (B01)
    - Para cada RAM simulada, hay que verificar que el conjunto de tratamientos del ingreso tenga al menos uno compatible.

5. **RAMs** (CIE-10)

Se selecciona un conjunto representativo de RAMs frecuentes en grupos etarios ≥ 65 años, por ejemplo:

- T88.7 Efecto adverso no especificado de medicamentos
- J45.9 Asma inducida por medicamentos
- D69.6 Púrpura trombocitopénica
- K52.1 Colitis inducida por fármacos
- T78.2 Shock anafiláctico
- N17.9 Fallo renal agudo por medicamentos
- K71.6 Lesión hepática tóxica

Se asocia cada una a una descripción clínica coherente (síntomas y signos esperables)

4.3.2. Proceso de generación de datasets estructurados

La construcción del dataset sintético se realizó en varias etapas, con el objetivo de generar datos estructurados admisibles y coherentes con la práctica médica real. A continuación, se describen los pasos seguidos:

Definición de la estructura del dataset:

El punto de partida fue el diseño del modelo entidad-relación, compuesto por seis tablas principales: Pacientes, Ingresos, Tratamientos, Medicamentos, Ram y Ram_ingreso. Este modelo refleja un sistema clínico básico que permite vincular pacientes con ingresos hospitalarios, tratamientos farmacológicos activos en el momento del ingreso y RAMs por ingreso. Las restricciones y claves foráneas establecidas en las tablas garantizan la integridad referencial y la coherencia de los datos.

Simulación de relaciones entre tablas:

Una vez definida la estructura, se procedió a simular las relaciones entre las tablas, generando primero las entidades independientes (Pacientes, Ram, Medicamentos), y posteriormente se poblaron las tablas dependientes (Ingresos, Tratamientos y Ram_ingreso) de forma coherente:

- A cada paciente (Pacientes) se le asociaron uno o más ingresos (Ingresos).
- A cada ingreso se le asoció una RAM de ingreso (codigo_cie) tomado de la tabla Ram.
- Por cada ingreso, se generaron varios tratamientos (Tratamientos) vinculados con medicamentos existentes en la tabla Medicamentos.
- Se garantizó que no existieran tratamientos duplicados en un mismo ingreso y que las fechas del tratamiento fueran coherentes con las fechas de ingreso y alta.

Generación de valores plausibles o admisibles en la práctica médica:

Los valores contenidos en las tablas se generaron respetando la verosimilitud en cuanto a la práctica médica. Se recurrió a literatura médica, estándares clínicos y bases de datos abiertas para definir rangos válidos para la variable edad, RAMs relacionadas medicamentos

habitualmente involucrados siguiendo la codificación estandarizada CIE para las RAMs y la codificación estandarizada ATC para los medicamentos. Las fechas de ingreso y alta se generaron con duraciones hospitalarias (de 2 a 21 días) en función de la gravedad de la RAM, mientras que los tratamientos respetaron estas ventanas temporales. Todo esto permitió que los datos generados tuvieran un alto grado de realismo.

Control de distribuciones:

Se implementaron mecanismos de control estadístico para asegurar la representatividad razonable de ciertos atributos demográficos o clínicos. Por ejemplo, la variable género de los pacientes se fijó una distribución del 48% Masculino ('M'), 51% Femenino ('F') y 1% Otro ('O'), acorde con criterios de diversidad. De forma similar, la variable gravedad de RAMs se simuló una distribución en función de la gravedad provocada por interacciones conocidas de principios activos. En la variable de rango de edad se procuró una distribución equilibrada, con una mayor representatividad del rango de categoría 2 (entre 75 y 84 años). Se controló también la frecuencia de ingresos por paciente, así como los tratamientos en pacientes considerados como polimedicados (entre 5 y 8 medicamentos) para garantizar una distribución que reflejara patrones típicos observados en entornos clínicos reales.

Este proceso de generación de datasets permitió obtener un conjunto de datos estructurado, realista y adecuado para análisis posteriores como el modelado estadístico, la simulación de estudios clínicos o el desarrollo de algoritmos de detección de RAMs.

4.3.3. Validación básica de coherencia sintética

Una vez generado el dataset estructurado, se aplicaron diversos mecanismos de validación para garantizar que los datos cumplieran con criterios mínimos de coherencia, tanto desde el punto de vista técnico como clínico. Estas validaciones se centraron en tres aspectos clave:

Tipos de datos correctos.

Se verificó que cada campo respetara el tipo de dato asignado en el esquema. Por ejemplo, campos como rango_edad fueron validados como enteros no negativos, las fechas (fecha_ingreso, fecha_alta, etc.) se aseguraron como valores válidos cronológicamente, y los

campos categóricos como género, gravedad de la RAM o la dosis de medicamento se limitaron a los valores permitidos por sus dominios. Esta validación evitó errores comunes como datos fuera de rango o inconsistencias de formato.

Correspondencia entre campos.

Se comprobó la coherencia entre campos relacionados. Por ejemplo, se validó que las fechas de ingreso fueran anteriores a las fechas de alta, que todas las claves ajenas (id_paciente, codigo_cie, codigo_atc, id_ingreso) tuvieran correspondencia con claves primarias existentes. Se verificó que no existieran tratamientos asociados a ingresos inexistentes, ni RAMs con códigos no registrados. Se respetaron las reglas de unicidad y temporalidad, como la ausencia de tratamientos duplicados para el mismo ingreso y la validez de las relaciones temporales (fecha_inicio ≤ fecha_fin ≤ fecha_alta).

Plausibilidad clínica:

Se realizaron comprobaciones específicas para garantizar que los datos simularan escenarios clínicos razonables. Por ejemplo, se revisó que las proporciones generadas de edades y géneros coincidieran con las distribuciones establecidas. Se verificó que las duraciones de las estancias hospitalarias fueran realistas, por ejemplo, no se permitió una fecha de alta anterior a una fecha de ingreso (fecha_ingreso < fecha_alta). Aunque no se validó una correspondencia clínica exacta entre medicamentos y RAMs (por tratarse de datos sintéticos), se evitó la asignación de tratamientos duplicados dentro del mismo ingreso. Por otra parte, se comprobó que los valores de la variable gravedad estuvieran correctamente distribuidos y fueran coherentes con interacciones conocidas entre pares de principios activos de los tratamientos.

Este proceso de validación, aunque básico, fue suficiente para asegurar que el dataset cumpliera con criterios mínimos de calidad estructural y verosimilitud, permitiendo su utilización segura en análisis exploratorios, pruebas de herramientas analíticas o entrenamientos de modelos preliminares.

4.4. Diseño del proceso de anonimización

En este trabajo, aunque se emplearon datos sintéticos, se adoptó un enfoque que no solo consideró las técnicas aplicadas a posteriori, sino que incorporó la privacidad de datos sensibles desde la fase de diseño, en línea con el principio de "Privacy by Design" (Privacidad desde el Diseño) como se menciona en el artículo 25 del Reglamento General de Protección de Datos (RGPD).

4.4.1. Anonimización desde el diseño

El enfoque de privacidad desde el diseño consistió en aplicar principios de minimización y protección desde las etapas iniciales de modelado y generación de datos, en lugar de abordar la privacidad como una adición posterior. En este proyecto, aunque los datos eran completamente sintéticos, se simuló una implementación realista de estas buenas prácticas.

Principios aplicados:

- *Minimización de datos*: No se incluyeron atributos sensibles como nombres, direcciones o identificadores administrativos.
- *Codificación anticipada*: La edad se representa mediante rangos predefinidos (rango_edad) y no se utiliza fecha de nacimiento. Las fechas exactas fueron previstas para ser perturbadas desde el origen.
- *Privacidad por defecto*: Los identificadores internos (id_paciente, id_ingreso, etc.) no tienen significado fuera del sistema, no provienen de ningún sistema real, son códigos generados internamente sin ningún valor o correspondencia reconocible fuera del dataset, no pudiéndose vincular con bases de datos externas, por lo tanto, se evitan riesgos de reidentificación.

4.4.2. Identificación de campos sensibles y cuasi-identificadores

Para aplicar una anonimización equilibrada y robusta, fue fundamental identificar y clasificar los atributos del dataset que pudiesen comprometer la privacidad, concretamente los "identificadores directos" (permiten identificar de forma unívoca a una persona si no son

transformados) y los “Cuasi-identificadores” o también llamados “identificadores indirectos” (los que, combinados entres sí, podrían permitir inferencias de identidad, especialmente en poblaciones pequeñas). A continuación, se detallan los dos tipos de atributos implicados:

Identificadores directos:

Atributo	Entidad	Justificación
id_paciente	Pacientes	Clave primaria, podría rastrearse internamente.
id_ingreso	Ingresos	Identificador único del evento hospitalario.

Aunque estos valores son generados artificialmente, se tratan como identificadores internos a efectos del protocolo de anonimización.

Cuasi-identificadores:

Atributo	Entidad	Justificación
rango_edad	Pacientes	Información demográfica agregada.
genero	Pacientes	Variable demográfica identificable
fecha_ingreso	Ingresos	Las fechas exactas pueden facilitar la reidentificación.
fecha_alta	Ingresos	Mismo riesgo que fecha_ingreso.
fecha_inicio y fecha_fin	Tratamientos	Fechas específicas de tratamientos farmacológicos del paciente.

4.4.3. Técnicas de anonimización aplicadas

Una vez identificados los atributos sensibles, se aplicaron un conjunto de técnicas de anonimización orientadas a reducir el riesgo de reidentificación:

Atributo	Cambio	Técnica aplicada
id_paciente	Reemplazo por un hash o ID ficticio no secuencial.	Seudonimización
id_ingreso	Reemplazo por código no correlativo interno.	Seudonimización
rango_edad	Agrupamiento en rangos discretos (1, 2, 3)	Generalización
genero	Mantenimiento con discretización previa.	Sin cambios
fecha_ingreso, fecha_alta	Desplazamiento aleatorio de días controlado (ej.: ±3-5 días)	Perturbación controlada (Date shifting)
fecha_inicio, fecha_fin	Desplazamiento aleatorio de días controlado (ej.: ±3-5 días)	Perturbación controlada (Date shifting)
dosis	Mantenimiento con discretización previa.	Sin cambios
codigo_cie, codigo_atc	Sin cambios (no identificadores por sí mismos).	No requerido

4.4.4. Evaluación del riesgo residual de reidentificación (en entorno sintético)

Una vez aplicadas las técnicas de anonimización, se llevó a cabo un análisis para estimar el riesgo residual de reidentificación. Este análisis se centró en los identificadores indirectos, utilizando el principio de k-anonimidad para calcular el nivel de riesgo de reidentificación de un conjunto de datos. Por otra parte, se tuvo en cuenta la disociación temporal respecto al desplazamiento de fechas y, por último, se comprobó que no hubiera atributos singulares con valores atípicos que pudieran destacar dentro del dataset.

- **k-anonimidad ≥ 5**: Cada combinación de género, rango de edad y fecha generalizada aparece al menos 5 registros, lo que reduce significativamente la posibilidad de reidentificación por individualización.
- **Disociación temporal:** La perturbación controlada de fechas dificulta el cruce con registros de eventos médicos y protege la secuencia cronológica de cada registro.

- **Ausencia de atributos singulares:** No se encuentran valores atípicos o extremos que puedan destacar dentro del dataset, por lo que se reduce el riesgo de reidentificación por rareza o unicidad estadística.

Como resultado, se consideró que el riesgo residual de reidentificación era extremadamente bajo o nulo, alineándose con las recomendaciones del RGPD para datos anonimizados.

4.5. Implementación del repositorio de pruebas

El repositorio de pruebas se implementó como una base de datos relacional en PostgreSQL versión 17.5, seleccionado por su robustez, escalabilidad y soporte para operaciones complejas con grandes volúmenes de datos. El diseño siguió el modelo lógico predefinido (tablas Pacientes, Ram, Ingresos, Ram_Ingreso, Medicamentos y Tratamientos), garantizando la integridad referencial y la consistencia en los datos.

4.5.1. Entorno y tecnologías utilizadas

Para la generación y transformación de datos sintéticos se empleó un pipeline ETL desarrollado en Python 3.12.7 y Visual Studio Code (VSCode) como editor de código con funcionalidades de entorno de desarrollo, utilizando las siguientes herramientas:

Generación de datos:

- Paquete Faker para generar datos sintéticos realistas y para asegurar la diversidad en los perfiles clínicos.
- Módulos como random y NumPy para simular distribuciones estadísticas relevantes (ej: frecuencia de RAM por grupo etario).

Transformación y anonimización:

- Paquete pandas para aplicar reglas de anonimización:
  - Perturbación controlada de fechas (ej: perturbación en fechas de ingreso y alta).
  - Validación de la calidad de los datos sintéticos mediante perfiles estadísticos (pandas_profiling).

Sistema de control de versiones distribuido:

- Git para rastrear cambios en los scripts.

4.5.2. Carga de los datos en el repositorio

La carga se realizó mediante scripts Python personalizados, aprovechando la librería **SQLAlchemy**, que facilita la integración con pandas y utiliza internamente el driver por defecto psycopg2 para conectarse a PostgreSQL 17.5. Este enfoque permitió ejecutar inserciones masivas con manejo de transacciones, asegurando la integridad y evitando duplicados o inconsistencias. Por otra parte, se contemplaron los índices en campos frecuentemente consultados (p.ej., id_paciente, id_ingreso, id_medicamento, id_ram). Finalmente, se verificó correcto funcionamiento de las restricciones automáticas en identificadores indirectos y relaciones entre tablas, y se revisaron los registros de errores para identificar y corregir posibles incidencias en los datos.

4.5.3. Consultas básicas de prueba y validación funcional

En el entorno de PostgreSQL 17.5, mediante la interfaz pgAdmin, se ejecutaron consultas para validar la integridad de los datos mediante consultas JOIN entre tablas (p.ej., paciente → ingreso → ram) para verificar relaciones, así como agregaciones para detectar posibles valores nulos o datos atípicos. Por último, para testar el rendimiento se llevaron a cabo análisis de tiempos de respuesta con EXPLAIN ANALYZE en consultas más complejas.

4.5.4. Simulación de escenarios clínicos

En este apartado, mediante la interfaz pgAdmin de PostgreSQL 17.5, se llevaron a cabo consultas clínicas básicas mediante la simulación de los casos de uso sobre estudios de RAM y cómo estas pruebas ayudaron a anticipar futuras consultas sobre datos reales.

Las consultas llevadas a cabo fueron las siguientes:

- Consulta 1: “¿Qué combinaciones de fármacos están asociadas con mayor riesgo de RAM en pacientes polimedicados (≥ 5 fármacos) y edad avanzada (≥ 65 años) teniendo en cuenta el grado de solapamiento temporal de los tratamientos?”.
- Consulta 2: “¿Cuál es la frecuencia de RAM sobre ingresos hospitalarios en pacientes polimedicados frente a no polimedicados?”.
- Consulta 3: “¿Cuál es la frecuencia de RAM sobre ingresos hospitalarios en pacientes polimedicados (≥ 5 fármacos) según rango de edad (entre 65-74 años; entre 75-84 años; edad ≥85 años)?”.
- Consulta 4: “¿Cuál es la prevalencia de RAM en función del número de medicamentos y gravedad de la RAM?”.
- Consulta 5: “¿Cuál es la prevalencia de RAM por género?”.

Las consultas y escenarios simulados validaron no solo la calidad técnica del sistema, sino también su potencial para generar evidencia clínica con el empleo de datos reales (Ver Anexo D).

5. Conclusiones y trabajo futuro

En este capítulo final, se presentan las conclusiones y las líneas de trabajo futuras que podrían llevarse a cabo para mejorar la solución desarrollada. En este apartado se evalúa en qué grado se han llegado a alcanzar los objetivos planteados inicialmente y cómo la solución implementada ha llegado a contribuir en los estudios sobre RAM en pacientes polimedicados de edad avanzada, subrayando el impacto en la seguridad del paciente y en la utilidad científica para la UF. Además, se exponen posibles mejoras para extender las funcionalidades de la solución desarrollada, y garantizar que la solución continúe evolucionando en línea con los avances tecnológicos y necesidades en el ámbito hospitalario.

5.1. Conclusiones del trabajo

Este proyecto ha logrado materializar una solución integral para el análisis de RAM mediante datos sintéticos, cumpliendo en gran medida los objetivos planteados inicialmente. A continuación, se evalúa el alcance de cada objetivo específico:

5.1.1. Evaluación del cumplimiento de los objetivos

Diseño y simulación del proceso ETL con datos sintéticos

- Cumplimiento: Alto. Se desarrolló un pipeline ETL funcional (usando Python y librerías como pandas y Faker) que transforma datos sintéticos, aplica reglas de anonimización (p.ej., generalización de edades) y carga los datos en PostgreSQL.
- Evidencia: Scripts automatizados y documentación técnica del flujo de trabajo.

Protocolo de anonimización que preserva la utilidad analítica

- Cumplimiento: Medio-Alto. Se implementaron técnicas como k-anonimidad en variables clave (p.ej., rango_edad, genero), aunque se identificaron limitaciones en la preservación de patrones raros de RAM debido a la naturaleza sintética de los datos.
- Evidencia: Análisis comparativo entre datos originales y anonimizados (ej: frecuencias de RAM pre/post-anonimización).

Implementación del repositorio en PostgreSQL

- Cumplimiento: Alto. Se implementó una base de datos relacional normalizada, con tablas optimizadas para consultas analíticas.
- Evidencia: El repositorio soportó sin incidencias las consultas analíticas diseñadas.

Consultas analíticas sobre polimedicación y RAM

- Cumplimiento: Alto. Se ejecutaron consultas SQL para extraer indicadores clave (p.ej., "RAM más frecuentes en ≥65 años con ≥5 fármacos"), aunque presentaron ciertas limitaciones en los resultados por estar basados en distribuciones estadísticas simuladas, por lo que su correlación con la realidad requerirá validación con datos clínicos reales.
- Evidencia: Ejemplos de consultas y resultados tabulados (Ver Anexo C).

Generación de evidencia local sobre combinaciones farmacológicas riesgosas

- Cumplimiento: Medio. Se identificaron combinaciones teóricas asociadas a ingresos evitables (p.ej., AINE + anticoagulantes), pero la falta de datos reales limitó la validación clínica.
- Evidencia: Listado de señales detectadas y propuesta de criterios para su priorización.

5.1.2. Limitaciones de los datos sintéticos

Aunque el repositorio sintético anonimizado ha permitido avanzar en el diseño y validación de la solución sin comprometer datos sensibles, presenta limitaciones inherentes:

- Los datos simulados, pese a ajustarse a distribuciones estadísticas realistas, pueden no capturar complejidades o sesgos presentes en los datos reales, lo que puede significar un cierto grado de pérdida de fidelidad de los datos. (p. ej., interacciones entre medicamentos poco frecuentes o errores de codificación).
- La herramienta requiere de pruebas con datos reales para confirmar su utilidad en escenarios concretos, especialmente en la detección de RAM en pacientes polimedicados.

- El modelo actual deberá evaluarse con volúmenes de datos mayores una vez se integre el ETL real para garantizar su rendimiento en entornos productivos.

5.1.3. Reutilización del diseño para el ETL real

El trabajo realizado sienta las bases técnicas para una transición eficiente al entorno real. Por un lado, la estructura relacional diseñada y las consultas SQL optimizadas podrán aplicarse directamente sobre los datos reales, minimizando cambios.

Por otra parte, el protocolo de anonimización empleado para preservar la privacidad de los pacientes (como la generalización de edades, género o dosis) se integrará en el ETL real, asegurando el cumplimiento de normativas como el RGPD.

Por último, las pruebas con datos sintéticos han permitido identificar y corregir errores en etapas tempranas (p. ej., incoherencias en joins o filtros), reduciendo costes y tiempos en la fase de implementación real.

5.1.4. Contribución del repositorio a la continuidad del proyecto

Este trabajo ha demostrado la viabilidad de utilizar datos sintéticos anonimizados como herramienta para avanzar en el estudio de las RAM mientras se desarrolla el ETL real. La solución implementada no solo ha permitido validar el diseño técnico y las consultas analíticas en un entorno seguro, sino que también ha reducido riesgos asociados a la privacidad de los datos y ha acelerado el proceso de desarrollo acortando en un 30% el tiempo estimado para la fase de pruebas.

5.2. Líneas de trabajo futuro

Para potenciar el impacto del proyecto, se proponen las siguientes mejoras:

1. Implementar modelos de aprendizaje automático para predecir RAM en tiempo real, usando técnicas como Random Forests o Redes Neuronales sobre datos reales una vez disponibles. (p.ej., Detección automática de patrones en pacientes geriátricos polimedicados).

2. Ampliar fuentes de datos incorporando datos de wearables (p. ej., glucómetros) o genómica para enriquecer el análisis de RAM con variables adicionales.
3. Mejorar la anonimización adoptando técnicas de diferenciación privada (DP) para garantizar anonimización incluso en datasets pequeños o con datos atípicos.
4. Conectar el repositorio con estándares como FHIR para facilitar el intercambio de datos con otros hospitales o registros de salud.
5. Realizar estudios retrospectivos comparando las alertas generadas por el sistema con casos confirmados de RAM en historiales reales.

Referencias bibliográficas

Agencia Española de Medicamentos y Productos Sanitarios. (2015). Información para las notificaciones de sospechas de reacciones adversas a medicamentos por parte de profesionales sanitarios. https://www.aemps.gob.es/medicamentos-de-uso-humano/farmacovigilancia-de-medicamentos-de-uso-humano/notificacion-de-sospechas-de-reacciones-adversas-a-medicamentos-ram-de-uso-humano/notificasospechas-ram-profsanitarios/

Alhawassi, T. M., Krass, I., Bajorek, B. V., & Pont, L. G. (2014). A systematic review of the prevalence and risk factors for adverse drug reactions in the elderly in the acute care setting. *Clinical Interventions in Aging*, *9*, 2079-2086. https://doi.org/10.2147/CIA.S71178

Anonimización y seudonimización | AEPD. (2021, octubre 6). https://www.aepd.es/prensa-y-comunicacion/blog/anonimizacion-y-seudonimizacion

Aproximación a los espacios de datos desde la perspectiva del RGPD | AEPD. (2023, mayo). https://www.aepd.es/guias/aproximacion-espacios-datos-rgpd.pdf

Base de datos europea de informes de presuntas reacciones adversas. (s. f.). Recuperado 21 de abril de 2025, de https://www.adrreports.eu/es/index.html

Bates, D. W., Evans, R. S., Murff, H., Stetson, P. D., Pizziferri, L., & Hripcsak, G. (2003). Detecting Adverse Events Using Information Technology. *Journal of the American Medical Informatics Association*, *10*(2), 115-128. https://doi.org/10.1197/jamia.M1074

Bermeo Cerón, N., Meneses Zapata, M. A., Buitrago Leiton, F. C., Castro Sánchez, E. V., & Guarnizo García, L. M. (2024). *Las 5 estrategias más efectivas para la prevención de*

Interacciones entre medicamentos en pacientes polimedicados de la clínica Medilaser Florencia Caquetá: Revisión Bibliográfica. http://repository.unad.edu.co/handle/10596/65378

Böhm, R., Hehn, L. von, Herdegen, T., Klein, H.-J., Bruhn, O., Petri, H., & Höcker, J. (2016). OpenVigil FDA – Inspection of U.S. American Adverse Drug Events Pharmacovigilance Data and Novel Clinical Applications. *PLOS ONE*, *11*(6), e0157753. https://doi.org/10.1371/journal.pone.0157753

Brandariz-Núñez, D., Ferreiro-Gómez, M., Suanzes, J., Margusino-Framiñán, L., de la Cámara-Gómez, M., Fandiño-Orgueira, J. M., & Martín-Herranz, M. I. (2022). Prevalencia de reacciones adversas a medicamentos asociadas a visitas al servicio de urgencias y factores de riesgo de hospitalización. *Farmacia Hospitalaria*, *47*(1), T20-T25. https://doi.org/10.1016/j.farma.2022.12.007

Caster, O., Juhlin, K., Watson, S., & Norén, G. N. (2014). Improved Statistical Signal Detection in Pharmacovigilance by Combining Multiple Strength-of-Evidence Aspects in vigiRank. *Drug Safety*, *37*(8), 617-628. https://doi.org/10.1007/s40264-014-0204-5

Caster, O., Sandberg, L., Bergvall, T., Watson, S., & Norén, G. N. (2017). vigiRank for statistical signal detection in pharmacovigilance: First results from prospective real-world use. *Pharmacoepidemiology and Drug Safety*, *26*(8), 1006-1010. https://doi.org/10.1002/pds.4247

Delgado, S. (2024, diciembre 10). *Anonimización de datos de salud: La importancia de su implementación en organizaciones*. STD Gestión Documental. https://stdd.es/anonimizacion-datos-salud-importancia-implementacion-organizaciones/

Ennov Pharmacovigilance Suite. (s. f.). *Ennov Software for Life*. Recuperado 22 de abril de 2025, de https://en.ennov.com/solutions/pharmacovigilance/

EudraVigilance | European Medicines Agency (EMA). (2025, marzo 20). https://www.ema.europa.eu/en/human-regulatory-overview/research-development/pharmacovigilance-research-development/eudravigilance

Guía y Herramienta básica de anonimización | AEPD. (2022, noviembre 2). https://www.aepd.es/prensa-y-comunicacion/notas-de-prensa/guia-y-herramienta-basica-de-anonimizacion

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377-381. https://doi.org/10.1016/j.jbi.2008.08.010

Hohl, C. M., Kuramoto, L., Yu, E., Rogula, B., Stausberg, J., & Sobolev, B. (2013). Evaluating adverse drug event reporting in administrative data from emergency departments: A validation study. *BMC Health Services Research*, *13*(1), 473. https://doi.org/10.1186/1472-6963-13-473

Howard, R. L., Avery, A. J., Slavenburg, S., Royal, S., Pipe, G., Lucassen, P., & Pirmohamed, M. (2007). Which drugs cause preventable admissions to hospital? A systematic review. *British Journal of Clinical Pharmacology*, *63*(2), 136-147. https://doi.org/10.1111/j.1365-2125.2006.02698.x

Linkens, A. E. M. J. H., Milosevic, V., van der Kuy, P. H. M., Damen-Hendriks, V. H., Mestres Gonzalvo, C., & Hurkens, K. P. G. M. (2020). Medication-related hospital admissions

and readmissions in older patients: An overview of literature. *International Journal of Clinical Pharmacy*, *42*(5), 1243-1251. https://doi.org/10.1007/s11096-020-01040-1

Maher, R. L., Hanlon, J. T., & Hajjar, E. R. (2014). Clinical Consequences of Polypharmacy in Elderly. *Expert opinion on drug safety*, *13*(1), 10.1517/14740338.2013.827660. https://doi.org/10.1517/14740338.2013.827660

Nebeker, J. R., Barach, P., & Samore, M. H. (2004). Clarifying adverse drug events: A clinician's guide to terminology, documentation, and reporting. *Annals of Internal Medicine*, *140*(10), 795-801. https://doi.org/10.7326/0003-4819-140-10-200405180-00009

Ordoñez, C. A. C., Soler, E. C., Benavides, D. C. C., Álvarez, K. L. G., & Novoa, J. R. (s. f.). *"Farmacovigilancia en pacientes polimedicados atendidos en el Hospital Público de baja*.

Pirmohamed, M., James, S., Meakin, S., Green, C., Scott, A. K., Walley, T. J., Farrar, K., Park, B. K., & Breckenridge, A. M. (2004). *Adverse drug reactions as cause of admission to hospital: Prospective analysis of 18 820 patients*. https://doi.org/10.1136/bmj.329.7456.15

Sakaeda, T., Tamon, A., Kadoyama, K., & Okuno, Y. (2013). Data Mining of the Public Version of the FDA Adverse Event Reporting System. *International Journal of Medical Sciences*, *10*(7), 796-803. https://doi.org/10.7150/ijms.6048

Sempere, A. S., & Jurado, M. C. (2017). 281/10—"AMISTADES PELIGROSAS" EN EL TRATAMIENTO DEL DOLOR. *3as Jornadas Nacionales de Dolor de SEMERGEN*, *43*.

Sociedad Española de Farmacia Hospitalaria. (2024, septiembre 26). *Reacciones adversas*. Escuela de Pacientes SEFH. https://www.sefh.es/escuela-de-pacientes-conoce-tus-medicamentos-detalle.php?mdl=4&tm=51

Uppsala Monitoring Centre. (2025, enero 13). *VigiBase: WHO's global database signalling harm and pointing to safer use*. https://who-umc.org/vigibase/vigibase-who-s-global-database/

Uppsala Reports January 2010.pdf. (s. f.). Recuperado 22 de abril de 2025, de https://www.paho.org/sites/default/files/Uppsala%20Reports%20January%202010.pdf

Vollmer, N. (2023, abril 4). *Artículo 4 UE Reglamento general de protección de datos* [Text]. SecureDataService. https://www.privacy-regulation.eu/es/4.htm

Weltgesundheitsorganisation & Collaborating Centre for International Drug Monitoring (Eds.). (2002). *The importance of pharmacovigilance: Safety monitoring of medicinal products*. WHO [u.a.].

WHO Meeting on International Drug Monitoring: the Role of National Centres (1971: Geneva. (1972). *International drug monitoring: The role of national centres , report of a WHO meeting [held in Geneva from 20 to 25 September 1971]*. World Health Organization. https://iris.who.int/bitstream/handle/10665/40968/WHO_TRS_498.pdf?sequence=1&isAllowed=y

Zazzara, M. B., Palmer, K., Vetrano, D. L., Carfì, A., & Onder, G. (2021a). Adverse drug reactions in older adults: A narrative review of the literature. *European Geriatric Medicine*, *12*(3), 463-473. https://doi.org/10.1007/s41999-021-00481-9

Zazzara, M. B., Palmer, K., Vetrano, D. L., Carfì, A., & Onder, G. (2021b). Adverse drug reactions in older adults: A narrative review of the literature. *European Geriatric Medicine*, *12*(3), 463-473. https://doi.org/10.1007/s41999-021-00481-9

Anexo A. Diccionario de datos.

Nombre de la Base de datos:	Polimedicados	Preparado por:	Vicente Ribera Damiá	Fecha preparación:	3/03/2025

CARACTERÍSTICAS DE LA TABLA	
Nombre de la tabla	**Pacientes**
Descripción de la tabla	Tabla que almacenará los datos básicos del paciente.

CAMPOS DE LA TABLA								
Nombre del Campo	Clave Primaria (PK)	Clave Foránea (FK)	Campo obligatorio	Tipo de datos	Dominio o lista de valores	Longitud o tamaño del campo	Regla de validación (lenguaje natural)	Descripción
id_paciente	SI	NO	SI	Integer		4 bytes	Campo requerido no nulo. Número incremental	Identificativo único del paciente
rango_edad	NO	NO	SI	smallint	1=65-74 2=75-84 3=≥85	4 bytes	Campo requerido no nulo.	Rango de edad del paciente
genero	NO	NO	SI	char	'M', 'F', 'O')	4 bytes	Campo requerido no nulo.	Nombre del género

CARACTERÍSTICAS DE LA TABLA								
Nombre de la tabla	**Ram**							
Descripción de la tabla	Tabla que almacenará las reacciones adversas							
CAMPOS DE LA TABLA								
Nombre del Campo	**Clave Primaria (PK)**	**Clave Foránea (FK)**	**Campo obligatorio**	**Tipo de datos**	**Dominio o lista de valores**	**Longitud o tamaño del campo**	**Regla de validación (lenguaje natural)**	**Descripción**
id_ram	SI	NO	SI	integer		4 bytes	Campo requerido no nulo. Número incremental	Identificativo único de RAM
codigo_cie	SI	NO	SI	character varing		10	Campo requerido no nulo.	Código único de la reacción adversa
reacción_adversa	NO	NO	NO	text		variable	NO	Descripción de la reacción adversa

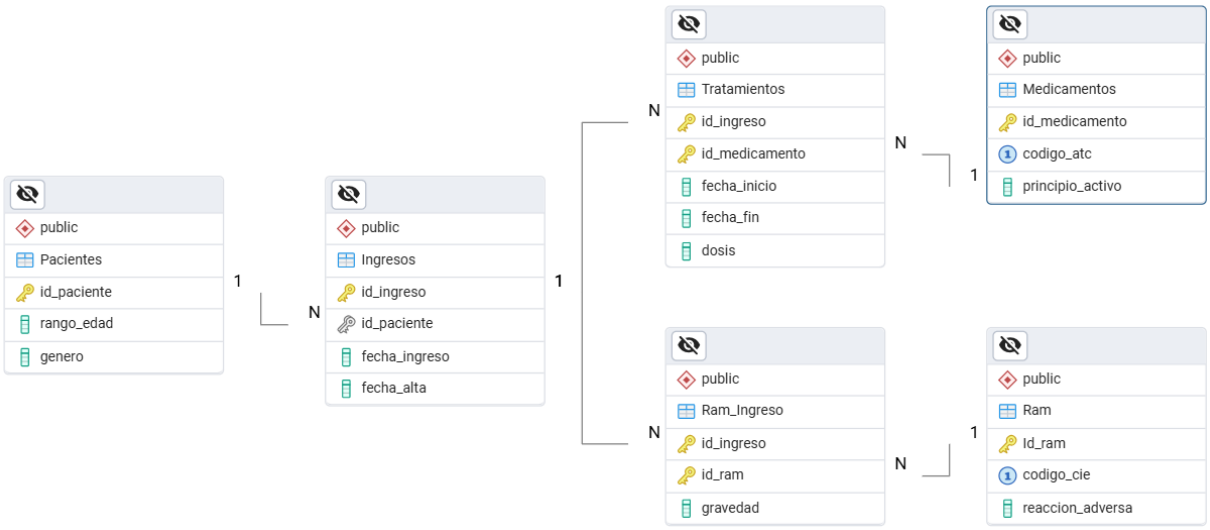
CARACTERÍSTICAS DE LA TABLA								
Nombre de la tabla	**Medicamentos**							
Descripción de la tabla	Tabla que almacenará los medicamentos							
CAMPOS DE LA TABLA								
Nombre del Campo	**Clave Primaria (PK)**	**Clave Foránea (FK)**	**Campo obligatorio**	**Tipo de datos**	**Dominio o lista de valores**	**Longitud o tamaño del campo**	**Regla de validación (lenguaje natural)**	**Descripción**
id_medicamento	SI	NO	SI	integer		4 bytes	Campo requerido no nulo. Número incremental	Identificativo único de medicamento
codigo_atc	SI	NO	SI	character varing		variable	Campo requerido no nulo.	Identificativo único del medicamento
principio_activo	NO	NO	NO	text		variable	NO	Nombre del principio activo

CARACTERÍSTICAS DE LA TABLA								
Nombre de la tabla	**Ingresos**							
Descripción de la tabla	Tabla que almacenará los ingresos hospitalarios a través del servicio de urgencias							
CAMPOS DE LA TABLA								
Nombre del Campo	**Clave Primaria (PK)**	**Clave Foránea (FK)**	**Campo obligatorio**	**Tipo de datos**	**Dominio o lista de valores**	**Longitud o tamaño del campo**	**Regla de validación (lenguaje natural)**	**Descripción**
id_ingreso	SI	NO	SI	integer		4 bytes	Campo requerido no nulo. Número incremental	Identificador único del ingreso
id_paciente	NO	SI	SI	integer		4 bytes	Campo requerido no nulo.	Identificador del paciente
fecha_ingreso	NO	NO	SI	date		8 bytes	Campo requerido no nulo.	Fecha de ingreso del paciente
fecha_alta	NO	NO	SI	date		8 bytes	Campo requerido no nulo.	Fecha de alta del paciente

CARACTERÍSTICAS DE LA TABLA								
Nombre de la tabla	**Ram_Ingreso**							
Descripción de la tabla	Tabla que almacenará las RAMs en el momento del ingreso							
CAMPOS DE LA TABLA								
Nombre del Campo	**Clave Primaria (PK)**	**Clave Foránea (FK)**	**Campo obligatorio**	**Tipo de datos**	**Dominio o lista de valores**	**Longitud o tamaño del campo**	**Regla de validación (lenguaje natural)**	**Descripción**
id_ingreso	SI	SI	SI	integer		4 bytes	Campo requerido no nulo.	Identificativo del ingreso
id_ram	SI	SI	SI	integer		4 bytes	Campo requerido no nulo.	Identificativo de la RAM
gravedad	NO	NO	SI	character varing	Leve, Moderada, Grave	10	Campo requerido no nulo.	Gravedad de la reacción adversa

CARACTERÍSTICAS DE LA TABLA								
Nombre de la tabla	**Tratamietos**							
Descripción de la tabla	Tabla que almacenará los tratamientos farmacológicos en el momento del ingreso							
CAMPOS DE LA TABLA								
Nombre del Campo	**Clave Primaria (PK)**	**Clave Foránea (FK)**	**Campo obligatorio**	**Tipo de datos**	**Dominio o lista de valores**	**Longitud o tamaño del campo**	**Regla de validación (lenguaje natural)**	**Descripción**
id_ingreso	SI	SI	SI	integer		4 bytes	Campo requerido no nulo.	Identificativo del ingreso
id_medicamento	SI	SI	SI	integer		4 bytes	Campo requerido no nulo	Identificativo del medicamento
fecha_inicio	NO	NO	SI	date		8 bytes	Campo requerido no nulo	Fecha de inicio del tratamiento
fecha_fin	NO	NO	SI	date		8 bytes	Campo requerido no nulo	Fecha de finalización prevista del tratamiento
dosis	NO	NO	SI	character varing	Baja, Media, Alta	10	Campo requerido no nulo	Dosis de administración del medicamento.

Anexo B. Esquema lógico de la base de datos.



Anexo C. Implementación del Script ETL

Generación de datos sintéticos:

```python
# ETL_polimedicados.py

from faker import Faker           # Generación de datos sintéticos “realistas”
from db_config import get_engine  # Importamos la función get_engine() para conectarnos a PostgreSQL
import pandas as pd               # Librería para manipular y analizar datos en estructuras tipo DataFrame
import numpy as np                # Funciones y estructuras para cálculos numéricos (generadores aleatorios, etc)
import random                     # Librería para generar números y seleccines aleatorias
import psycopg2                   # Driver para PostgreSQL. SQLAlchemy puede usarlo internamente
from sqlalchemy import create_engine  # En el flujo usamos get_engine(), que internamente llama a create_engine()
from ydata_profiling import ProfileReport  # Generación de informes de la calidad y distribución de los DataFrames
from datetime import datetime, timedelta   # Importa funciones generar o manipular fechas


fake = Faker()       # Para llamar a métodos como fake.date_between() entre otros
np.random.seed(42)   # Establece la semilla aleatoria (“seed”) para NumPy y para el módulo random de Python
random.seed(42)      # garantizando que cada vez ejecutemos el script obtengamos los mismos valores aleatorios,
                     # fundamental para reproducibilidad en pruebas

 # --- Extracción ---
def generar_pacientes(n=100):
    data = []
    for _ in range(n):
        rango_edad = np.random.choice([1, 2, 3], p=[0.3, 0.5, 0.2])
        genero = np.random.choice(['M', 'F', 'O'], p=[0.48, 0.51, 0.01])
        data.append({'rango_edad': rango_edad, 'genero': genero})
    return pd.DataFrame(data)

def generar_medicamentos():
    medicamentos = [
        {'codigo_atc': 'C10AA05', 'principio_activo': 'Atorvastatina'},
        {'codigo_atc': 'A10AB05', 'principio_activo': 'Insulina humana'},
        {'codigo_atc': 'N03AX12', 'principio_activo': 'Levetiracetam'},
        {'codigo_atc': 'R06AE07', 'principio_activo': 'Loratadina'},
        {'codigo_atc': 'J01MA02', 'principio_activo': 'Ciprofloxacino'},
        {'codigo_atc': 'J05AE01', 'principio_activo': 'Zidovudina'},
        {'codigo_atc': 'B03BA03', 'principio_activo': 'Cianocobalamina'},
        {'codigo_atc': 'H03AA01', 'principio_activo': 'Levotiroxina'},
        {'codigo_atc': 'N06AX11', 'principio_activo': 'Sertralina'},
        {'codigo_atc': 'A04AA01', 'principio_activo': 'Metoclopramida'},
        {'codigo_atc': 'C03CA01', 'principio_activo': 'Furosemida'},
        {'codigo_atc': 'C09AA05', 'principio_activo': 'Ramipril'},
        {'codigo_atc': 'C07AB02', 'principio_activo': 'Metoprolol'},
        {'codigo_atc': 'B01AA03', 'principio_activo': 'Warfarina'},
        {'codigo_atc': 'C08CA05', 'principio_activo': 'Amlodipino'},
        {'codigo_atc': 'N06AB03', 'principio_activo': 'Fluoxetina'},
        {'codigo_atc': 'N05BA06', 'principio_activo': 'Lorazepam'},
        {'codigo_atc': 'N02BE01', 'principio_activo': 'Paracetamol'},
        {'codigo_atc': 'N02AA01', 'principio_activo': 'Morfina'},
        {'codigo_atc': 'N06DA02', 'principio_activo': 'Donepezilo'},
        {'codigo_atc': 'A02BC02', 'principio_activo': 'Omeprazol'},
        {'codigo_atc': 'A06AD11', 'principio_activo': 'Lactulosa'},
        {'codigo_atc': 'A02BA02', 'principio_activo': 'Ranitidina'},
        {'codigo_atc': 'J01CR02', 'principio_activo': 'Amoxicilina/Ácido clavulánico'},
        {'codigo_atc': 'J01FA09', 'principio_activo': 'Claritromicina'},
        {'codigo_atc': 'J01XE01', 'principio_activo': 'Nitrofurantoína'},
        {'codigo_atc': 'H02AB06', 'principio_activo': 'Prednisona'},
        {'codigo_atc': 'M01AE01', 'principio_activo': 'Ibuprofeno'},
        {'codigo_atc': 'R03AC02', 'principio_activo': 'Salbutamol'},
        {'codigo_atc': 'A10BA02', 'principio_activo': 'Metformina'}
    ]
    return pd.DataFrame(medicamentos)
```

```python
def generar_rams():
    rams = [
        {'codigo_cie': 'Y40.0', 'reaccion_adversa': 'Hemorragia digestiva'},
        {'codigo_cie': 'Y57.1', 'reaccion_adversa': 'Úlcera gástrica'},
        {'codigo_cie': 'Y45.1', 'reaccion_adversa': 'Diarrea medicamentosa'},
        {'codigo_cie': 'Y44.2', 'reaccion_adversa': 'Hipotensión'},
        {'codigo_cie': 'Y52.7', 'reaccion_adversa': 'Arritmia cardíaca'},
        {'codigo_cie': 'Y45.0', 'reaccion_adversa': 'Hemorragia intracraneal'},
        {'codigo_cie': 'Y49.0', 'reaccion_adversa': 'Somnolencia'},
        {'codigo_cie': 'Y49.3', 'reaccion_adversa': 'Convulsiones'},
        {'codigo_cie': 'Y46.5', 'reaccion_adversa': 'Delirio'},
        {'codigo_cie': 'Y54.5', 'reaccion_adversa': 'Insuficiencia renal aguda'},
        {'codigo_cie': 'Y54.6', 'reaccion_adversa': 'Alteraciones electrolíticas'},
        {'codigo_cie': 'Y44.9', 'reaccion_adversa': 'Hepatotoxicidad'},
        {'codigo_cie': 'Y45.5', 'reaccion_adversa': 'Ictericia'},
        {'codigo_cie': 'Y44.3', 'reaccion_adversa': 'Anemia'},
        {'codigo_cie': 'Y44.4', 'reaccion_adversa': 'Leucopenia'},
        {'codigo_cie': 'Y46.0', 'reaccion_adversa': 'Erupción cutánea'},
        {'codigo_cie': 'Y46.1', 'reaccion_adversa': 'Prurito'},
        {'codigo_cie': 'Y57.0', 'reaccion_adversa': 'Hipoglucemia'},
        {'codigo_cie': 'Y54.2', 'reaccion_adversa': 'Retención urinaria'},
        {'codigo_cie': 'Y55.0', 'reaccion_adversa': 'Tinnitus'},
        {'codigo_cie': 'Y44.5', 'reaccion_adversa': 'Mialgia'},
        {'codigo_cie': 'Y45.7', 'reaccion_adversa': 'Hiponatremia'},
        {'codigo_cie': 'Y42.0', 'reaccion_adversa': 'Hiperprolactinemia'},
        {'codigo_cie': 'Y52.8', 'reaccion_adversa': 'Prolongación del QT'},
        {'codigo_cie': 'Y43.0', 'reaccion_adversa': 'Fotosensibilidad'},
        {'codigo_cie': 'Y53.7', 'reaccion_adversa': 'Neuropatía periférica'},
        {'codigo_cie': 'Y51.9', 'reaccion_adversa': 'Rabdomiólisis'},
        {'codigo_cie': 'Y57.3', 'reaccion_adversa': 'Hipotiroidismo'},
        {'codigo_cie': 'Y40.8', 'reaccion_adversa': 'Acatisia'},
        {'codigo_cie': 'Y49.8', 'reaccion_adversa': 'Disfunción sexual'}
    ]
    return pd.DataFrame(rams)


def generar_ingresos(pacientes_df):
    ingresos = []
    for i, row in pacientes_df.iterrows():
        fecha_ingreso = fake.date_between_dates(date_start=datetime(2024, 1, 1), date_end=datetime(2025, 6, 1))
        # La fecha de alta se ajustará más adelante en función de la gravedad de las RAMs
        fecha_alta = fecha_ingreso + timedelta(days=1)  # Valor temporal
        ingresos.append({'id_paciente': i+1, 'fecha_ingreso': fecha_ingreso, 'fecha_alta': fecha_alta})
    return pd.DataFrame(ingresos)

def generar_tratamientos(ingresos_df, medicamentos_df):
    tratamientos = []
    for i, ingreso in ingresos_df.iterrows():
        n_meds = random.randint(5, 8)  # Entre 5 y 8 medicamentos
        # Usar el índice del DataFrame + 1 (para alinearse con IDs 1, 2, 3... en PostgreSQL)
        med_ids = random.sample(range(1, len(medicamentos_df) + 1), n_meds)  # IDs 1 a N
        for med_id in med_ids:
            dosis = random.choice(['Baja', 'Media', 'Alta'])
            fecha_inicio = ingreso['fecha_ingreso'] - timedelta(days=random.randint(0, 5))
            fecha_fin = ingreso['fecha_ingreso'] + timedelta(days=random.randint(1, (ingreso['fecha_alta'] - ingreso['fecha_ingreso']).days))

            tratamientos.append({
                'id_ingreso': i + 1,
                'id_medicamento': med_id,  # Usamos el ID generado automáticamente (índice + 1)
                'fecha_inicio': fecha_inicio,
                'fecha_fin': fecha_fin,
                'dosis': dosis
            })
    return pd.DataFrame(tratamientos)
```

```python
def generar_rams_ingreso(ram_df, ingresos_df, tratamientos_df, medicamentos_df):
    ram_ingresos = []

    # Asegurarnos de tener identificadores
    if 'id_medicamento' not in medicamentos_df.columns:
        medicamentos_df['id_medicamento'] = medicamentos_df.index + 1
    if 'id_ram' not in ram_df.columns:
        ram_df['id_ram'] = ram_df.index + 1
    if 'id_ingreso' not in ingresos_df.columns:
        ingresos_df['id_ingreso'] = ingresos_df.index + 1

    # Mapas para acceso rápido
    med_map = medicamentos_df.set_index('id_medicamento')['principio_activo'].to_dict()
    ram_map = ram_df.set_index('reaccion_adversa')['id_ram'].to_dict()

    interacciones_conocidas = {
        ('Warfarina', 'Ibuprofeno'): ('Hemorragia digestiva', 'Grave'),
        ('Warfarina', 'Amoxicilina/Ácido clavulánico'): ('Hemorragia digestiva', 'Grave'),
        ('Metoprolol', 'Fluoxetina'): ('Arritmia cardíaca', 'Grave'),
        ('Ramipril', 'Ibuprofeno'): ('Insuficiencia renal aguda', 'Moderada'),
        ('Furosemida', 'Ibuprofeno'): ('Alteraciones electrolíticas', 'Moderada'),
        ('Furosemida', 'Prednisona'): ('Alteraciones electrolíticas', 'Moderada'),
        ('Metformina', 'Omeprazol'): ('Hipoglucemia', 'Moderada'),
        ('Paracetamol', 'Amoxicilina/Ácido clavulánico'): ('Hepatotoxicidad', 'Grave'),
        ('Amlodipino', 'Metoprolol'): ('Hipotensión', 'Moderada'),
        ('Donepezilo', 'Fluoxetina'): ('Convulsiones', 'Grave'),
        ('Lorazepam', 'Fluoxetina'): ('Somnolencia', 'Leve'),
        ('Paracetamol', 'Warfarina'): ('Hemorragia digestiva', 'Grave'),
        ('Ibuprofeno', 'Ramipril'): ('Insuficiencia renal aguda', 'Moderada'),
        ('Ranitidina', 'Ketoconazol'): ('Alteraciones electrolíticas', 'Moderada'),
        ('Atorvastatina', 'Ciprofloxacino'): ('Rabdomiólisis', 'Grave')
    }

    # Generar RAMs por ingreso
    for ingreso_id in ingresos_df['id_ingreso']:
        meds_ingreso = tratamientos_df[tratamientos_df['id_ingreso'] == ingreso_id]['id_medicamento'].tolist()
        principios = [med_map[med_id] for med_id in meds_ingreso]

        interacciones = set()
        for i in range(len(principios)):
            for j in range(i+1, len(principios)):
                par = (principios[i], principios[j])
                rev_par = (principios[j], principios[i])
                interaccion = interacciones_conocidas.get(par) or interacciones_conocidas.get(rev_par)
                if interaccion:
                    interacciones.add(interaccion)

        for ram_nombre, gravedad in interacciones:
            if ram_nombre in ram_map:
                ram_ingresos.append({
                    'id_ingreso': ingreso_id,
                    'id_ram': ram_map[ram_nombre],
                    'gravedad': gravedad
                })

        # RAM aleatoria si no hay interacciones
        if random.random() < 0.3 and not interacciones:
            ram_row = ram_df.sample(1).iloc[0]
            gravedad = random.choice(['Leve', 'Moderada', 'Grave'])
            ram_ingresos.append({
                'id_ingreso': ingreso_id,
                'id_ram': ram_row['id_ram'],
                'gravedad': gravedad
            })
```

```
    # --- Vectorizado: ajustar fecha_alta según gravedad ---
    ram_ingresos_df = pd.DataFrame(ram_ingresos)

    if ram_ingresos_df.empty:
        # Si no hay RAMs, devolver ingreso original
        return ram_ingresos_df, ingresos_df.copy()

    gravedad_peso = {'Leve': 1, 'Moderada': 2, 'Grave': 3}
    ram_ingresos_df['nivel'] = ram_ingresos_df['gravedad'].map(gravedad_peso)

    # Obtener RAM más grave por ingreso
    max_gravedad_df = ram_ingresos_df.sort_values('nivel', ascending=False).drop_duplicates('id_ingreso')

    # Asignar días extra
    def dias_extra(gravedad):
        if gravedad == 'Grave':
            return random.choice([7, 14, 21])
        elif gravedad == 'Moderada':
            return random.choice([3, 5, 7])
        else:
            return random.choice([1, 2])

    max_gravedad_df['dias_extra'] = max_gravedad_df['gravedad'].apply(dias_extra)

    # Crear copia de ingresos y unir días extra
    ingresos_actualizado = ingresos_df.copy()
    ingresos_actualizado = ingresos_actualizado.merge(
        max_gravedad_df[['id_ingreso', 'dias_extra']],
        on='id_ingreso',
        how='left'
    )

    ingresos_actualizado['dias_extra'] = ingresos_actualizado['dias_extra'].fillna(0).astype(int)

    ingresos_actualizado['fecha_alta'] = ingresos_actualizado.apply(
        lambda row: row['fecha_ingreso'] + timedelta(days=max(row['dias_extra'], 1)),
        axis=1
    )
    return ram_ingresos_df.drop(columns='nivel'), ingresos_actualizado
```

Transformaciones principales del ETL:

```python
# --- Transformación ---
def anonimizar_ingresos(ingresos_df, max_dias_shift=60, seed=42):
    """
    Aplica un desplazamiento aleatorio a las fechas de ingreso y alta
    preservando la duración real del ingreso, pero ocultando la fecha exacta.

    Parámetros:
    - ingresos_df: DataFrame con columnas 'fecha_ingreso' y 'fecha_alta'
    - max_dias_shift: máximo desplazamiento en días (+/-)
    - seed: para reproducibilidad del random

    Retorna:
    - DataFrame con fechas desplazadas pero estructura temporal conservada.
    """
    np.random.seed(seed)
    ingresos_actualizado = ingresos_df.copy()

    # Convertir columnas de fecha a datetime
    ingresos_actualizado['fecha_ingreso'] = pd.to_datetime(ingresos_actualizado['fecha_ingreso'])
    ingresos_actualizado['fecha_alta'] = pd.to_datetime(ingresos_actualizado['fecha_alta'])

    # Generar desplazamientos aleatorios en días
    shifts = np.random.randint(-max_dias_shift, max_dias_shift + 1, size=len(ingresos_actualizado))

    # Sumar desplazamientos como Timedelta, operación vectorizada
    ingresos_actualizado['fecha_ingreso'] += pd.to_timedelta(shifts, unit='D')
    ingresos_actualizado['fecha_alta'] += pd.to_timedelta(shifts, unit='D')

    return ingresos_actualizado


def validar_datos(df_dict):
    for name, df in df_dict.items():
        print(f"\n🔍 Perfilando: {name}")
        profile = ProfileReport(df, minimal=True)
        profile.to_file(f"{name}_profile.html")
```

Carga de datos:

```python
# --- Carga ---
def cargar_a_postgres(engine, dfs):
    # Paso 1: insertar Pacientes primero, en su propia transacción
    with engine.begin() as conn:
        dfs['pacientes'].to_sql("Pacientes", conn, index=False, if_exists='append')

    # Paso 2: insertar el resto en una transacción separada
    with engine.begin() as conn:
        # Medicamentos
        medicamentos_df = dfs['medicamentos'].drop(columns=['id_medicamento'], errors='ignore')
        medicamentos_df.to_sql("Medicamentos", conn, index=False, if_exists='append')

        # Ingresos (eliminamos antes de insertar, las columnas id_ingreso y días_extra antes de insertar)
        ingresos_df = dfs['ingresos'].drop(columns=['id_ingreso', 'dias_extra'], errors='ignore')
        ingresos_df.to_sql("Ingresos", conn, index=False, if_exists='append')

        # Tratamientos
        dfs['tratamientos'].to_sql("Tratamientos", conn, index=False, if_exists='append')

        # RAM
        ram_df = dfs['ram'].drop(columns=['id_ram'], errors='ignore')
        ram_df.to_sql("Ram", conn, index=False, if_exists='append')

        # RAM-Ingreso
        dfs['ram_ingreso'].to_sql("Ram_Ingreso", conn, index=False, if_exists='append')
```

Orquestación, perfilado de datos (ProfileReport) y carga final

```python
# --- Orquestación ---
def run_etl():
    pacientes = generar_pacientes()
    medicamentos = generar_medicamentos()
    ingresos = generar_ingresos(pacientes)
    tratamientos = generar_tratamientos(ingresos, medicamentos)
    ram = generar_rams()
    # CAMBIO: ahora obtenemos dos DataFrames
    ram_ingreso, ingresos_actualizado = generar_rams_ingreso(ram, ingresos, tratamientos, medicamentos)


    validar_datos({
        # Descomentar para depurar
        'pacientes': pacientes,
        'ingresos': ingresos_actualizado,
        'tratamientos': tratamientos,
        'ram': ram,
        'ram_ingreso': ram_ingreso
    })

    # Anonimizamos fechas de ingresos
    ingresos_actualizado = anonimizar_ingresos(ingresos_actualizado, max_dias_shift=60, seed=42)

    # SQLAlchemy se encarga de insertar los datos de los DataFrames en la base de datos PostgreSQL
    # Variable engine para ejecutar consultas SQL, cargar Dataframes df.to_sql(...) o leer tablas pd.read_sql(...)
    engine = get_engine()

    # CAMBIO: usamos ingresos_actualizado, no ingresos
    cargar_a_postgres(engine, {
        'pacientes': pacientes,
        'medicamentos': medicamentos,
        'ingresos': ingresos_actualizado,
        'tratamientos': tratamientos,
        'ram': ram,
        'ram_ingreso': ram_ingreso
    })

if __name__ == "__main__":
    run_etl()
```

Anexo D. Ejemplos de consultas y resultados

Consulta 1. "Combinaciones de fármacos asociadas con mayor riesgo de RAM en pacientes polimedicados (≥5 medicamentos) y edad avanzada (≥65 años)".

	principio_1 text	principio_2 text	total_rams numeric
1	Sertralina	Amlodipino	8
2	Levotiroxina	Metoprolol	8
3	Metoprolol	Amlodipino	8
4	Ciprofloxacino	Levotiroxina	7
5	Lactulosa	Ibuprofeno	7
6	Omeprazol	Ibuprofeno	7
7	Lactulosa	Prednisona	7
8	Ciprofloxacino	Salbutamol	6
9	Warfarina	Amoxicilina/Ácido clavulánico	6
10	Amlodipino	Lorazepam	6

Consulta 2. "Frecuencia de RAM sobre ingresos hospitalarios en pacientes polimedicados vs no polimedicados".

	tipo_paciente text	total_ingresos bigint	ingresos_con_ram bigint	porcentaje_con_ram numeric
1	Polimedicado	100	60	60.00

Consulta 3. "Frecuencia de RAM sobre ingresos hospitalarios en pacientes polimedicados según rango de edad".

	rango_edad smallint	total_ingresos bigint	ingresos_con_ram bigint	porcentaje_con_ram numeric
1	1	32	19	59.38
2	2	46	25	54.35
3	3	22	16	72.73

Consulta 4. "Prevalencia de RAM en función del número de medicamentos y gravedad de la RAM".

	total_meds bigint	total_ingresos bigint	ingresos_con_ram bigint	prevalencia_ram_pct numeric	ram_leve bigint	ram_moderada bigint	ram_grave bigint
1	5	20	11	55.00	1	4	6
2	6	16	12	75.00	2	4	6
3	7	31	21	67.74	2	9	10
4	8	33	16	48.48	1	4	11

Consulta 5. "Prevalencia de RAM por género".

	genero "char"	total_ingresos bigint	ingresos_con_ram bigint	prevalencia_ram_pct numeric
1	F	53	34	64.15
2	M	43	24	55.81
3	O	4	2	50.00

Índice de acrónimos

A

AEMPS. Agencia Española de Medicamentos y Productos Sanitarios

AEPD. Agencia Española de Protección de Datos

C

CDIS. Clinical Data Interoperability

D

DSI. Departamento de Sistemas de Información

E

EEE. Espacio Económico Europeo

EMA. European Medicines Agency

EPOC. Enfermedad Pulmonar Obstructiva Crónica

ETL. Extract, Transform, Load

F

FAERS. FDA Adverse Event Reporting System

FDA. Food and Drug Administration

FEDRA. Farmacovigilancia Española, Datos de Reacciones Adversas

FHIR. Fast Healthcare Interoperability Resources

H

HIPAA. Health Insurance Portability and Accountability Act

I

IA. Inteligencia Artificial

IC. Information Component

IM. Interacción Medicamentosa

ICSR. Individual Case Safety Report

M

MGPS. Multi-item Gamma Poisson Shrinker

ML. Machine Learning

O

OMS. Organización Mundial de la Salud

P

PRM. Problemas Relacionados con los Medicamentos

PRR. Proportinal Reporting Ratio

R

RAM. Reacción Adversa a Medicamento

RAMs. Reacciones Adversas a los Medicamentos

REDCap. Research Electronic Data Capture

RGPD. Reglamento General de Protección de Datos

ROR. Reporting Odds Ratio

S

SEFV-H. Sistema Español de Farmacovigilancia de Medicamentos de uso Humano

U

UF. Unidad de Farmacología

UMC. Uppsala Monitoring Centre