Detección de Señales en FAERS: Ciencia de Datos en Farmacovigilancia

1. Introducción y Motivación

La farmacovigilancia analiza eventos adversos asociados al uso de medicamentos después de su autorización de comercialización. El sistema *FAERS* (FDA Adverse Event Reporting System) reúne millones de reportes espontáneos, con información demográfica del paciente, sospechas de medicamentos y reacciones adversas codificadas (p.ej., MedDRA).

Este proyecto propone un **pipeline reproducible** para detectar y priorizar *señales* de seguridad entendidas como *combinaciones fármaco–evento* cuya frecuencia relativa excede la esperada bajo independencia. Los objetivos prácticos son: (i) monitoreo de seguridad post-comercialización, (ii) estratificación por demografía, y (iii) análisis temporal de persistencia o emergencia de señales, con salidas interpretables y auditables.

Aportes esperados. Un conjunto de herramientas y artefactos reproducibles (código, datos derivados, figuras y tablero) que puedan servir de base a un manuscrito corto de métodos/aplicación en farmacovigilancia.

2. Objetivos

2.1. Objetivo General

Desarrollar un pipeline reproducible de ciencia de datos para **detectar y priorizar señales de seguridad** en FAERS mediante medidas de desproporcionalidad y análisis temporal, con salidas interpretables (tablas y visualizaciones interactivas).

2.2. Objetivos Específicos

- Implementar la **ingesta y limpieza** de los extractos trimestrales de FAERS (deduplicación por `caseid`/`version`).
- Estandarizar nombres de sustancias (*ingrediente activo*) y codificación de eventos (MedDRA).
- Calcular medidas de desproporcionalidad: **PRR**, **ROR** e **IC** (opcional: **EBGM** con *Empirical Bayes*).

- Construir **series temporales trimestrales** por combinación fármaco–evento e incorporar **estratificación** (sexo, edad, país, tipo de reportante).
- Desarrollar un **tablero** con ranking de señales, filtros y *drill-down* por indicación y demografía.
- Documentar **buenas prácticas** de reproducibilidad (versionado, *data dictionary*, pruebas unitarias) y **evaluación** (sanity checks, sensibilidad).

3. Marco Teórico

3.1. Señales en sistemas de reporte espontáneo

Un reporte espontáneo es una notificación de una sospecha de reacción adversa tras el uso de un medicamento. Para una pareja (D, A) (medicamento D, evento adverso A), se define una tabla 2×2 a partir de coocurrencias en la base:

	Evento A	No A
Medicamento D	n_{11}	n_{10}
Otros meds	n_{01}	n_{00}

donde $N = n_{11} + n_{10} + n_{01} + n_{00}$.

3.2. Medidas de desproporcionalidad

Las medidas comparan el *riesgo/odds* de reportar A cuando se reporta D contra el resto de la base.

PRR (Proportional Reporting Ratio).

$$\mathrm{PRR} = \frac{\frac{n_{11}}{n_{11}+n_{10}}}{\frac{n_{01}}{n_{01}+n_{00}}}. \tag{1}$$

ROR (Reporting Odds Ratio).

$$\mathrm{ROR} = \frac{n_{11}/n_{10}}{n_{01}/n_{00}} = \frac{n_{11}n_{00}}{n_{10}n_{01}}. \tag{2}$$

Intervalos de confianza. Para log(ROR) es común aproximar

$$\mathrm{SE}[\log(\mathrm{ROR})] \approx \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}}}, \tag{3}$$

y construir IC del 95 % como $\log(\mathrm{ROR}) \pm 1{,}96\,\mathrm{SE}$. Se aplica **corrección de Haldane** $(+\frac{1}{2})$ cuando algún conteo es cero.

IC (Information Component, BCPNN).

$$\mathrm{IC} = \log_2\left(\frac{n_{11}\,N}{(n_{11}+n_{10})(n_{11}+n_{01})}\right), \qquad \mathrm{Var}(\mathrm{IC}) \approx \frac{1}{(\ln 2)^2}\left(\frac{1}{n_{11}} - \frac{1}{n_{11}+n_{10}} - \frac{1}{n_{11}+n_{01}} + \frac{1}{N}\right). \tag{4}$$

EBGM (opcional). La *Empirical Bayes Geometric Mean* incorpora *shrinkage* hacia la media global, estabilizando razones con conteos bajos. Requiere ajuste de hiperparámetros (p.ej., mediante *Gamma-Poisson*).

3.3. Ajustes metodológicos clave

- **Deduplicación:** mantener la última `version` por `caseid`; excluir duplicados históricos.
- **Rol del fármaco:** diferenciar `PS` (*Primary Suspect*), `SS` (*Secondary Suspect*), `C` (concomitante). Análisis principal: `PS`.
- **Normalización:** mapear a ingrediente activo; estandarizar eventos a términos MedDRA preferidos.
- **Sesgos:** *under-reporting*, *notoriety* y cambios regulatorios; mitigación vía estratificación y análisis temporal.
- **Temporalidad:** cortes trimestrales; *nowcasting* para reportes tardíos; detección de cambio (CUSUM/estadísticos de *scan*).

4. Metodología

4.1. Datos y Esquema FAERS

Fuentes. Extractos trimestrales (ASCII/XML) con tablas principales: `DEMO` (demografía), `DRUG` (medicamentos), `REAC` (eventos), `INDI` (indicaciones).

Llaves. `caseid` identifica el caso; `primaryid` identifica el reporte; `version` permite deduplicación temporal.

Variables mínimas. Sexo, edad (estandarizada a años), país, rol del fármaco, fecha de recepción, término MedDRA, nombre del fármaco.

4.2. Implementación Algorítmica

1. **Ingesta:** descarga y lectura de archivos; conversión a un *data lake* (`parquet`/`duckdb`).
2. **Limpieza:** deduplicación por `caseid`/`version`; filtrado `PS`; estandarización de unidades/edades.
3. **Unión y conteos:** enlace `DRUG`–`REAC` por `primaryid`; construcción de (D, A) y totales del resto (n_{10}, n_{01}, n_{00}).

4. **Desproporcionalidad:** cálculo de PRR, ROR, IC; IC95 % mediante aproximaciones normales; Haldane si hay ceros.
5. **Temporal:** series trimestrales por (D, A); suavizado opcional (media móvil); **tendencias** (Mann–Kendall) y **rupturas** (Bai–Perron).
6. **Priorización:** reglas (umbral en PRR ≥ 2, ROR IC95 % > 1, #casos ≥ 3 y *persistencia* ≥ 2 trimestres); ranking por severidad (eventos `serious`).
7. **Visualización:** barras top-k, *sparklines* temporales, *forest plots* por estrato; tabla interactiva con filtros (sustancia, evento, sexo, edad).
8. **Validación cualitativa:** revisión manual de casos; comparación con etiquetado de seguridad (trabajo futuro, ver §9).

4.3. Evaluación del Desempeño

- **Sanity checks:** replicar señales canónicas (p.ej., eventos con *Boxed Warning*).
- **Sensibilidades:** variar umbrales; analizar `PS` vs `SS`; estratos demográficos.
- **Eficiencia:** tiempo de ejecución por trimestre/año; uso de memoria; escalamiento a 5–10 años.

4.4. Reproducibilidad y Calidad

- **Control de versiones** (git) y **entornos** (conda/poetry).
- **Data dictionary** y **esquemas** (tablas base y derivadas).
- **Pruebas** unitarias para funciones clave (deduplicación, conteos, IC).
- **Registros** (*logs*) y **metadatos** (fecha de descarga, versión FAERS).

5. Resultados Esperados

- **Dataset** limpio/documentado con series trimestrales de PRR/ROR/IC por (D, A).
- **Tablas** Top-20 por clase terapéutica, con #casos, IC95 % y persistencia.
- **Figuras** principales: barras top-k; evolución temporal de señales; *forest plots* por estrato.
- **Tablero interactivo** (Dash/Streamlit) para exploración.
- **Informe técnico** reproducible (notebook + `Makefile`/`pyproject`).

6. Herramientas y Tecnologías

- **Lenguaje:** Python (`pandas`, `numpy`, `scipy`, `statsmodels`).
- **Visualización:** `matplotlib`, `plotly`.

- **Almacenamiento:** `duckdb`/`sqlite`; archivos `parquet`.
- **Opcional:** `lifelines` (temporal), `dash`/`streamlit` (tablero).

7. Consideraciones Éticas y Limitaciones

- **No inferir causalidad**: FAERS es un sistema pasivo con sesgos; las señales son *hipótesis* que requieren evaluación clínica/regulatoria.
- **Privacidad**: trabajar sólo con datos públicos y anonimados; no intentar reidentificar.
- **Sesgos de reporte**: variación en calidad/completitud; cambios por notoriedad o regulación.

8. Plan de Trabajo y Cronograma

1. **Semana 1–2:** Ingesta, exploración y **deduplicación** (script base, pruebas).
2. **Semana 3–4:** Construcción de (D, A), conteos y **PRR/ROR/IC** con IC95 %.
3. **Semana 5: Series temporales**, tendencias y rupturas.
4. **Semana 6: Priorización** y **tablero** mínimo viable.
5. **Semana 7:** Validación, sensibilidades y **documentación**.

9. Trabajo Futuro y Publicación

- **Vincular con etiquetado (SPL/DailyMed):** minería de texto para detectar cuándo aparecen nuevas advertencias y si siguen picos en FAERS.
- **Modelado Bayesiano (EBGM):** comparación sistemática con PRR/ROR/IC para señales raras.
- **Manuscrito corto**: sección de métodos y caso de estudio por clase terapéutica; discusión de implicaciones regulatorias.

10. Bibliografía breve

Literatura base de desproporcionalidad (PRR/ROR/IC) y documentación técnica de FAERS/MedDRA/FDA. (Se ampliará con referencias específicas en la versión final del informe.)