Generación de registros electrónicos de salud sintéticos a partir de una cohorte de >1 millón de pacientes diabéticos andaluces

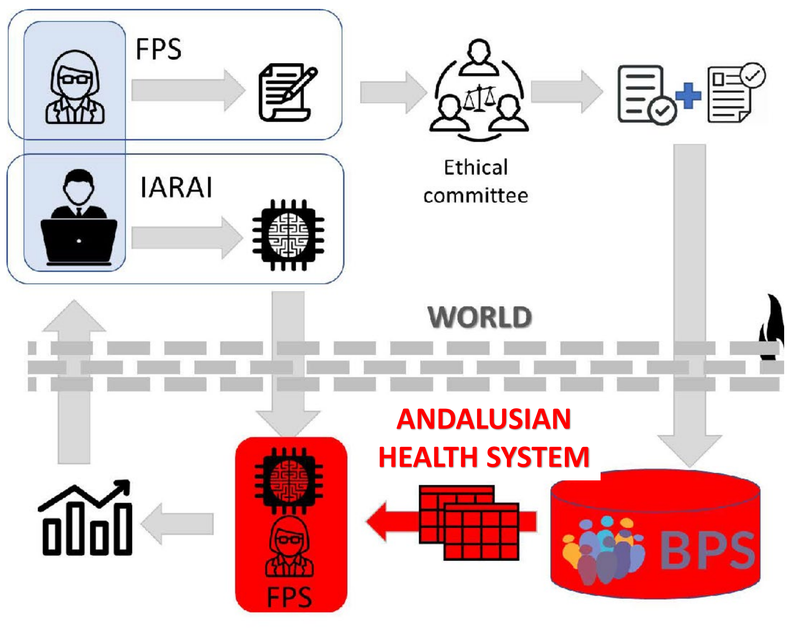
Francisco M. Ortuño[1,2], Carlos Loucera[2,3], Laura Alejos[2], David Kreil[4], Joaquin Dopazo[2,3,5,6]

[1] Departamento de Ingeniería de Computadores, Automática y Robótica, Universidad de Granada (UGR), Granada, España.
[2] Plataforma de Medicina Computacional, Fundación Progreso y Salud (FPS), Sevilla, España.
[3] Medicina Computacional de Sistemas, Instituto de Biomedicina de Sevilla (IBiS), Sevilla, España.
[4] *Institute of Advanced Research in Artificial Intelligence* (IARAI), Vienna, Austria.
[5] Bioinformática en Enfermedades Raras (BiER), Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Seville, Spain.
[6] FPS/ELIXIR-ES, Virgen del Rocío Hospital, Seville, Spain.

OBJECTIVO

La **Base Poblacional de Salud (BPS)** es un recurso del Sistema Andaluz de Salud en el que se recogen de forma curada y estructurada registros electrónicos de salud de **>15 millones de pacientes**, con información clínica de los últimos 20 años. Sin embargo, debido a las restricciones por la Ley de Protección de Datos y la privacidad de los datos clínicos, la explotación de este recurso para el análisis secundario de datos clínicos y su posible utilización como una potente fuente de datos y evidencias del mundo real (RWD/RWE) son muy limitadas. Por ello, este trabajo propone la generación de datos clínicos sintéticos realistas obtenidos a partir de la información extraída de la BPS, utilizando varios modelos generativos computacionales como los *Generative Adversarial Networks* (GANs), *Variational AutoEncoders* (VAEs) *o Diffussion Models* (DMs). Estos modelos consiguen generar información que mantiene las características, asociaciones y patrones intrínsecos en los datos reales, sin exponer datos privados y evitando posibles re-identificaciones. Los modelos propuestos se han utilizado para generar y validar un dataset clínico sintético de diagnósticos obtenido a partir de una cohorte real de **1.215.8974 pacientes diabéticos.**

METODOLOGÍA

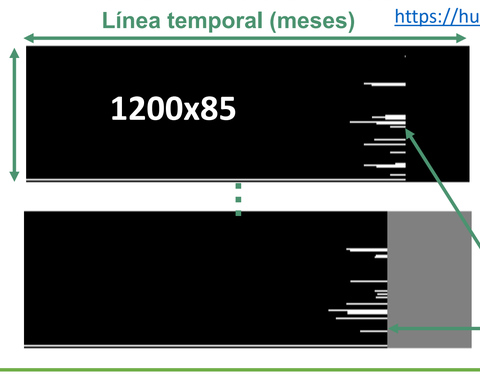


	Original	Curados
Pacientes	1.215.974	979.308
Visitas médicas	6.174.049	5.854.580
Códigos grabados	7.752.446	6.003.164
Códigos Patologías	82	83* + sexo

** Añadido procedimiento de amputación*

Modelo Difusión Incondicional (UDM)

Línea temporal (meses) https://huggingface.co/docs/diffusers/index

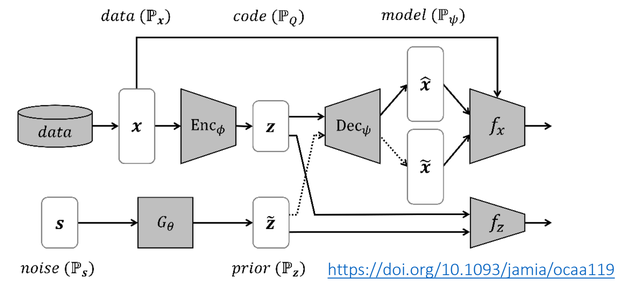


Modelo DDPM UNet
batch size = 8
50 épocas
4 up/down layers +
1 up/down attention

Pérdida de censado o fallecimiento

Modelo dualAAE

- Combinación de autoencoder con GANs.
- Generación de una sucesión ordenada de visitas médicas con diagnósticos.



https://doi.org/10.1093/jamia/ocaa119

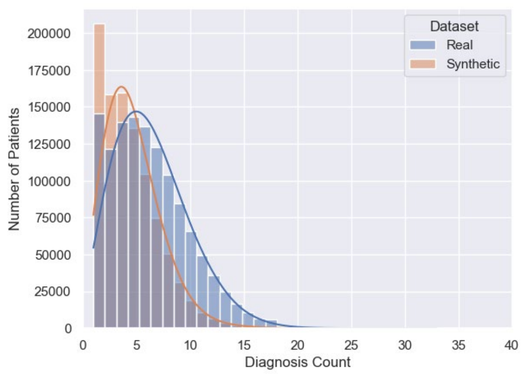
- Eliminados diagnósticos sin fecha o edad
- Eliminados amputados previo diabetes
- Eliminados diabéticos diagnosticados <2003

RESULTADOS

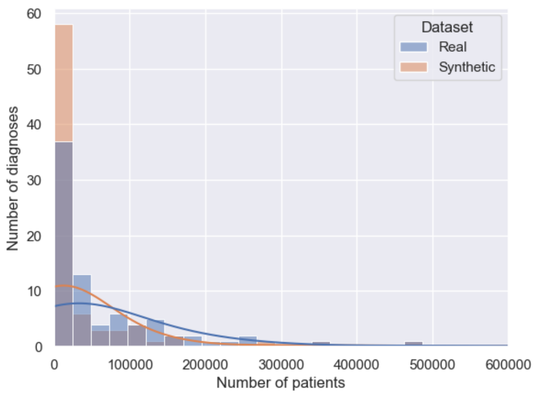
DualAAE: 999.936 Pacientes Diabéticos Sintéticos

dualAAE

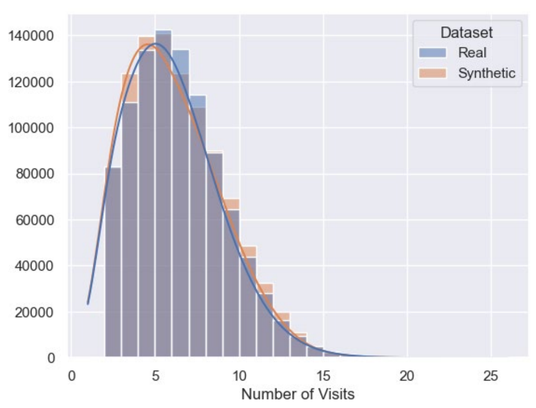
Diagnósticos por paciente



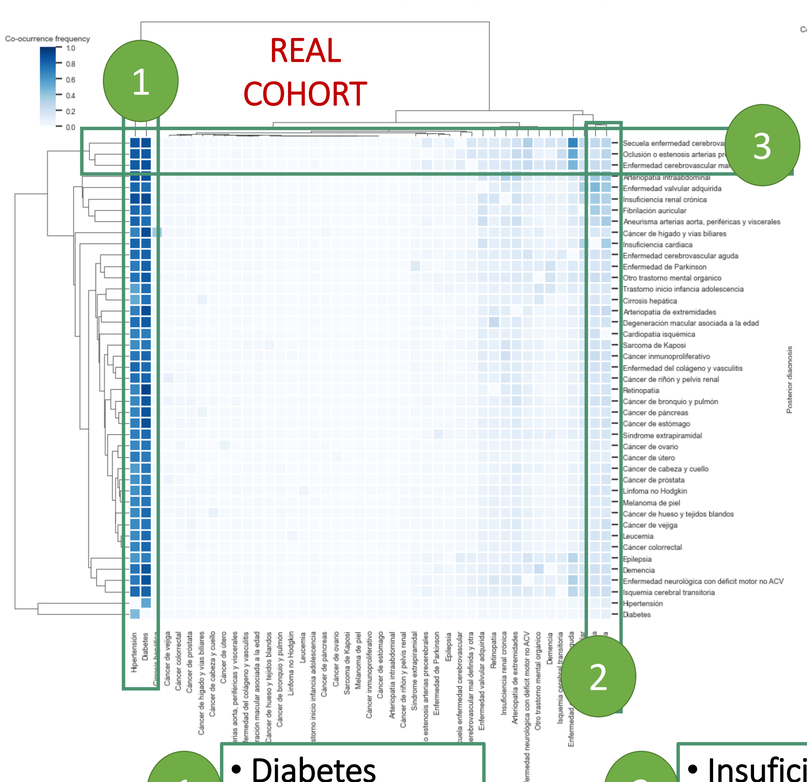
Pacientes por diagnóstico

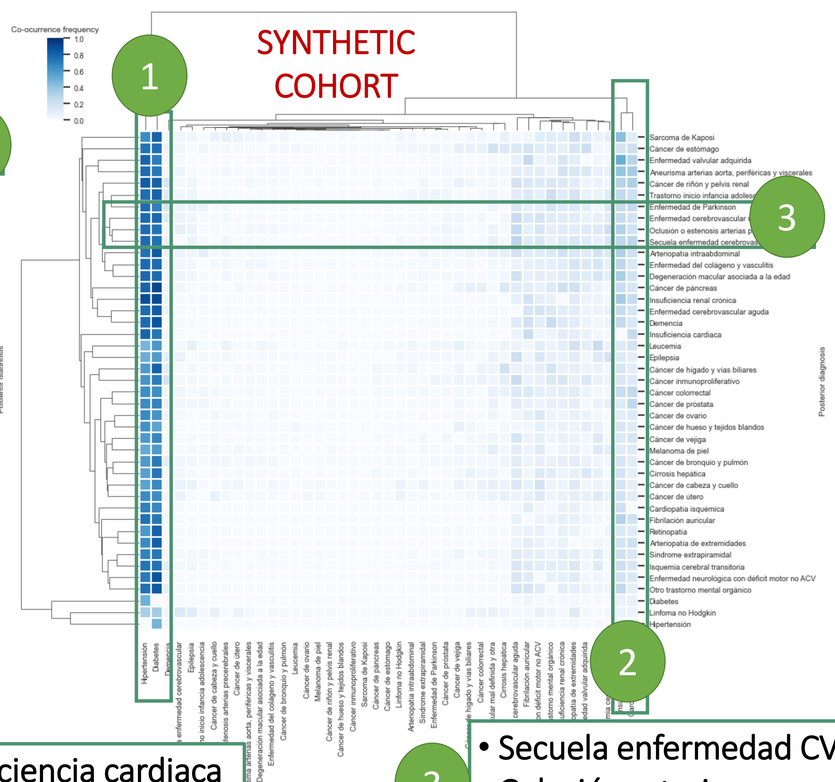


Visitas por paciente



Secuencias de co-ocurrencias más frecuentes



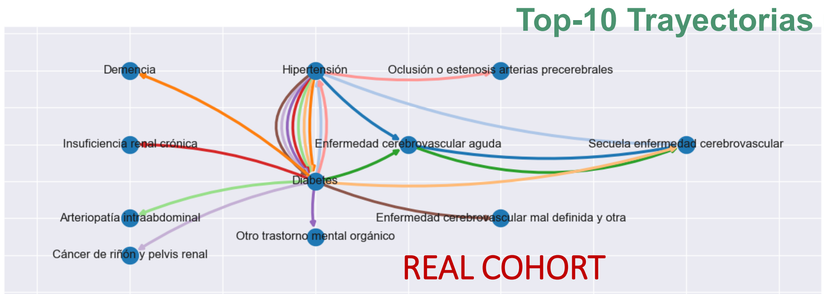


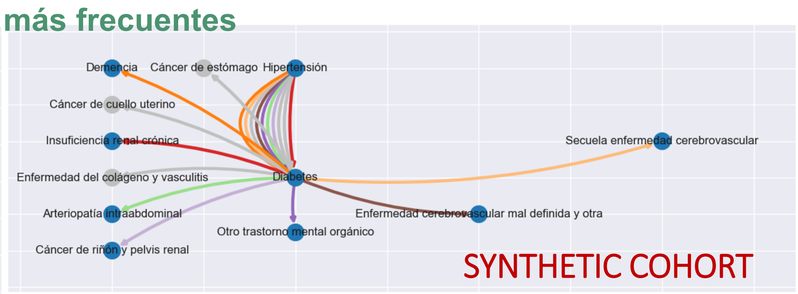
1 • Diabetes
• Hypertension

2 • Insuficiencia cardiaca
• Cardiopatía isquémica

3 • Secuela enfermedad CV
• Oclusión arterias precerebrales
• Enfermedad CV mal definida

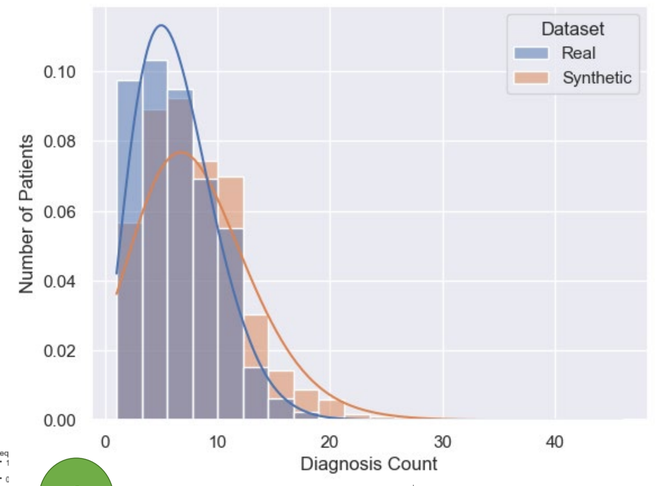
Top-10 Trayectorias más frecuentes



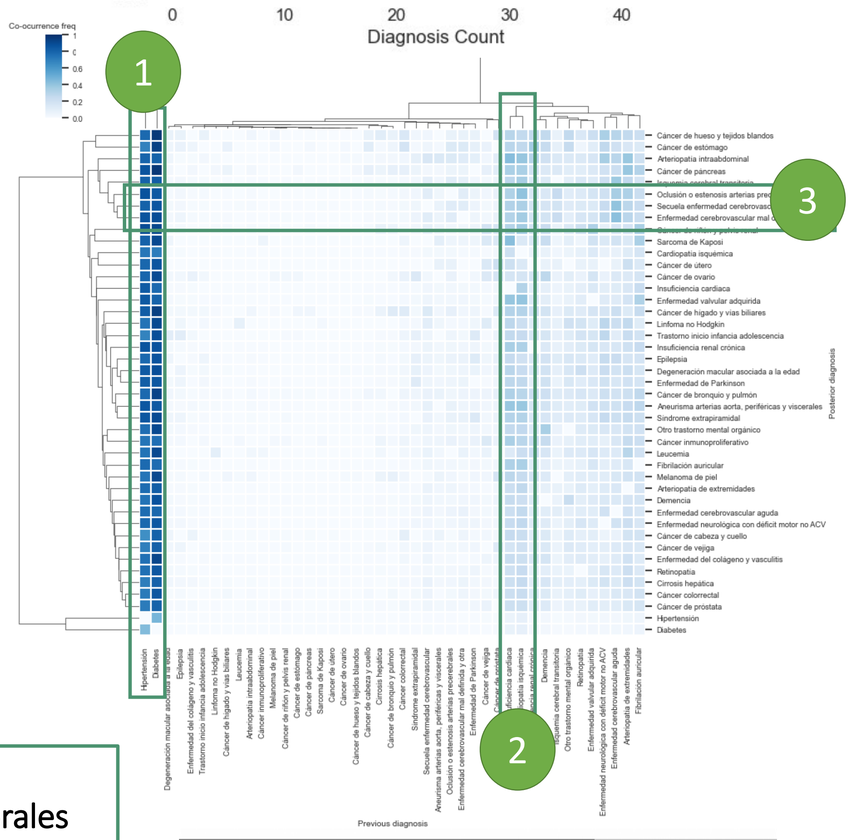


UDM: 9.316 Imágenes eHR Sintéticas

Diagnósticos por paciente



Secuencias de co-ocurrencias más frecuentes



DEFECTOS CORREGIDOS / ELIMINADOS:
- Pacientes sin diabetes
- Diagnósticos intermitentes o repetidos
- Imágenes o visitas vacías

CONCLUSION

Los modelos conocidos basados en **GANs consiguen generar eHRs realistas basados en secuencias de diagnósticos**, simulando patrones de enfermedad y co-ocurrencias reales. Sin embargo, estos modelos están limitados ya que **no pueden modelar la temporalidad de las variables clínicas registradas**. Por ello, se ha comenzado a estudiar la posibilidad de incorporar modelos basados en la imagen (difusión) que nos permiten controlar mejor esta temporalidad.

Contact Information: fortuno@ugr.es @clinicalbioinfo

Affiliations: Andalusian Platform for Computational Medicine · INB Instituto Nacional de Bioinformática · Universitas Granatensis · IBiS Instituto de Biomedicina de Sevilla · IARAI