Synthetic Data for Secure Biomedical Research in Privacy-Restricted Contexts: An Evaluatable Utility and Privacy Methodology

[1] Instituto de Investigaciones en Microbiología (IIM), Facultad de Ciencias, Universidad Nacional Autónoma de Honduras (UNAH), Tegucigalpa, 11101, Honduras. *liliansosa2012@gmail.com*. https://orcid.org/0000-0001-8148-9482

[2] Grupo de Investigación en Tecnología Farmacéutica, Facultad de Ciencias Químicas y Farmacia, Universidad Nacional Autónoma de Honduras (UNAH), Tegucigalpa, 11101, Honduras. *liliansosa2012@gmail.com*. https://orcid.org/0000-0001-8148-9482

[3] Departamento de Química, Universidad Técnica Particular de Loja (UTPL), Loja 110107, Ecuador. *lcespinoza@utpl.edu.ec* https://orcid.org/0000-0003-1486-5622

[4] Departamento de Química, Universidad Nacional Autónoma de Honduras (UNAH), Tegucigalpa, 11101, Honduras. *fredy.rodriguez@unah.edu.hn* . https://orcid.org/0000-0002-6516-1011

Abstract– ***Biomedical research increasingly relies on large clinical and epidemiological datasets, but sharing this information is often limited by privacy concerns, legal regulations, ethical requirements, and institutional barriers. This paper presents a reproducible methodology for generating and evaluating synthetic biomedical tabular data in privacy-restricted environments. The proposed framework combines data governance, real-data profiling, synthetic data generation, utility evaluation, and privacy-risk assessment before any data release. A proof of concept was developed using the publicly available Wisconsin Diagnostic Breast Cancer dataset. This dataset was selected as a controlled and auditable test case; because it is small, complete, and entirely numeric, it represents a favorable scenario for the generator rather than the heterogeneous records that motivate this work. A transparent Gaussian copula model generated a synthetic training dataset with the same size as the original training partition. Utility was evaluated through statistical fidelity, preservation of correlation structure, and train-synthetic-test-real predictive performance. Privacy was assessed using nearest-neighbor distance, exact-copy detection, and distance-based membership inference analysis. The results showed high statistical fidelity, minimal loss of predictive performance, no exact copies of real records, and membership inference performance close to random guessing. These findings show that the evaluation workflow is operational in a controlled setting and that evaluated synthetic data can support safer biomedical research collaboration without assuming that synthetic datasets are automatically anonymous or free of privacy risk. Transfer to heterogeneous clinical data under real privacy restrictions, as well as comparison with alternative generators, requires further validation.***

Keywords-- ***synthetic data, biomedical privacy, health data sharing, machine learning, privacy-preserving analytics.***

Datos sintéticos para investigación biomédica segura en contextos con restricciones de privacidad: metodología evaluable de utilidad y privacidad

[1] Instituto de Investigaciones en Microbiología (IIM), Facultad de Ciencias, Universidad Nacional Autónoma de Honduras (UNAH), Tegucigalpa, 11101, Honduras. *liliansosa2012@gmail.com*. https://orcid.org/0000-0001-8148-9482

[2] Grupo de Investigación en Tecnología Farmacéutica, Facultad de Ciencias Químicas y Farmacia, Universidad Nacional Autónoma de Honduras (UNAH), Tegucigalpa, 11101, Honduras. *liliansosa2012@gmail.com*. https://orcid.org/0000-0001-8148-9482

[3] Departamento de Química, Universidad Técnica Particular de Loja (UTPL), Loja 110107, Ecuador. *lcespinoza@utpl.edu.ec* https://orcid.org/0000-0003-1486-5622

[4] Departamento de Química, Universidad Nacional Autónoma de Honduras (UNAH), Tegucigalpa, 11101, Honduras. *fredy.rodriguez@unah.edu.hn* . https://orcid.org/0000-0002-6516-1011

Resumen: ***La investigación biomédica depende cada vez más de grandes conjuntos de datos clínicos y epidemiológicos, pero compartir esta información suele verse limitado por preocupaciones de privacidad, regulaciones legales, requisitos éticos y barreras institucionales. Este artículo presenta una metodología reproducible para generar y evaluar datos biomédicos sintéticos tabulares en entornos con restricciones de privacidad. El marco propuesto combina la gobernanza de datos, el perfilado de datos reales, la generación de datos sintéticos, la evaluación de la utilidad y la evaluación de los riesgos de privacidad antes de cualquier publicación de datos. Se desarrolló una prueba de concepto utilizando el conjunto de datos público de diagnóstico de cáncer de mama de Wisconsin. Este conjunto se eligió como caso controlado y auditable; al ser pequeño, completo y exclusivamente numérico, representa un escenario favorable para el generador y no reproduce la heterogeneidad de los registros clínicos que motivan el estudio. Un modelo de cópula gaussiana transparente generó un conjunto de datos de entrenamiento sintético del mismo tamaño que la partición de entrenamiento original. La utilidad se evaluó mediante fidelidad estadística, preservación de la estructura de correlación y rendimiento predictivo de entrenamiento-sintético-prueba-real. La privacidad se evaluó mediante la distancia al vecino más cercano, la detección de copias exactas y el análisis de inferencia de pertenencia basado en la distancia. Los resultados mostraron alta fidelidad estadística, mínima pérdida de rendimiento predictivo, ausencia de copias exactas de registros reales y un rendimiento de inferencia de pertenencia cercano al de un clasificador aleatorio.. Estos hallazgos muestran que el flujo de evaluación es operable en un entorno controlado y que los datos sintéticos evaluados pueden contribuir a una colaboración más segura en la investigación biomédica, sin asumir que los conjuntos de datos sintéticos sean automáticamente anónimos ni estén libres de riesgos para la privacidad. Su transferencia a datos clínicos heterogéneos con restricciones reales de privacidad y su comparación con generadores alternativos requieren validación adicional.***

Palabras clave: ***datos sintéticos, privacidad biomédica, intercambio de datos de salud, aprendizaje automático, análisis que preservan la privacidad.***

I. Introducción

El intercambio de datos biomédicos es una condición habilitante para la investigación traslacional, la vigilancia epidemiológica, el desarrollo de modelos predictivos y la innovación tecnológica en salud [1], [2]. Sin embargo, los datos biomédicos contienen patrones clínicos, demográficos, genéticos, administrativos y temporales que pueden permitir la identificación directa o indirecta de personas, aun cuando se hayan eliminado nombres, números de expediente o documentos de identidad [3]-[7]. Esta tensión entre el valor científico y la protección individual limita la colaboración interinstitucional, retrasa la validación externa de los algoritmos y reduce la posibilidad de construir ecosistemas de innovación basados en datos [8].

La literatura reciente identifica los datos sintéticos como una alternativa para ampliar el acceso a datos de salud con fines de investigación, prueba de algoritmos, educación, desarrollo de sistemas y la liberación pública de conjuntos de datos con distintos grados de utilidad [9]. En salud, la reutilización de datos exige cumplir con marcos normativos, estándares de seguridad, revisión ética y acuerdos de uso. En el ámbito de la Health Insurance Portability and Accountability Act (HIPAA), la guía del U.S. Department of Health and Human Services reconoce dos rutas para la desidentificación de la información protegida de salud: la

determinación experta y el método de puerto seguro [10]. En el marco europeo, el Reglamento General de Protección de Datos distingue entre datos personales, datos seudonimizados y datos anónimos y reserva la exclusión del régimen de protección únicamente para la información que no se relacione con una persona identificada o identificable [11]. Por tanto, producir un conjunto sintético no equivale automáticamente a anonimizarlo; se requiere evidencia cuantitativa de su utilidad y de su privacidad.

Este trabajo aborda la dificultad de compartir datos sensibles en investigación biomédica mediante una solución metodológica: generar datos sintéticos evaluables y someterlos a una batería mínima de pruebas antes de su uso o de su liberación. La propuesta se ubica en el cruce de la transformación digital, la inteligencia artificial y la gestión tecnológica, porque ofrece un procedimiento replicable para hospitales, universidades, centros de investigación, unidades de innovación y entidades públicas que requieren compartir evidencia sin exponer datos personales. El objetivo no es sustituir la gobernanza de datos ni relajar los requisitos éticos, sino proveer una capa técnica verificable que habilite la colaboración segura.

Las contribuciones del artículo son tres. ***Primero,*** se propone un flujo metodológico de generación sintética y de evaluación con métricas de fidelidad, utilidad predictiva y privacidad. Segundo, se presenta un esquema operativo de decisión que vincula las métricas con acciones de liberación, restricción o regeneración del conjunto sintético. Tercero, se implementa una prueba de concepto reproducible con un conjunto biomédico público, lo que permite mostrar resultados cuantitativos sin utilizar información clínica sensible. El alcance de esta prueba es deliberadamente acotado: su propósito es demostrar que el flujo de evaluación es operable y auditable, no establecer el desempeño comparativo de un generador ni su generalización a datos clínicos heterogéneos.

II. Fundamentos conceptuales y trabajo relacionado

A. Privacidad, desidentificación y riesgo residual

La desidentificación reduce el riesgo de identificación, pero no elimina de manera absoluta la posibilidad de reidentificación cuando existen variables raras, combinaciones únicas o fuentes externas enlazables. El National Institute of Standards and Technology (NIST) describe la desidentificación como un conjunto de aproximaciones orientadas a reducir el riesgo de compartir información y preservar una utilidad específica para usos posteriores [12]. Esta formulación es particularmente relevante para datos sintéticos, porque la utilidad y la privacidad no son atributos independientes: un conjunto demasiado parecido al real puede conservar utilidad, pero también aumentar el riesgo de memorizar registros; un conjunto excesivamente perturbado puede proteger mejor, pero perder valor analítico.

La privacidad diferencial constituye una garantía formal que limita el aumento del riesgo individual asociado a participar en una base de datos [13]. No todos los métodos de generación sintética son diferencialmente privados; por ello, cuando no se entrena bajo un mecanismo formal de privacidad diferencial, debe reportarse con claridad que la privacidad es empírica o diagnóstica, no garantizada matemáticamente. Además, los modelos generativos y predictivos pueden filtrar información a partir de los registros de entrenamiento. Los ataques de inferencia de membresía buscan determinar si un registro específico participó en el entrenamiento de un modelo, lo que puede revelar información sensible en contextos clínicos [14].

B. Generación de datos sintéticos tabulares

Los modelos generativos modernos permiten aprender distribuciones complejas y muestrear registros artificiales que preservan los patrones estadísticos del dato original. Las redes generativas adversarias fueron propuestas como un marco para entrenar modelos generativos mediante un juego minimax entre un generador y un discriminador [15]. En datos tabulares, la heterogeneidad de variables continuas, categóricas, multimodales o desequilibradas hace que la generación de datos sea más compleja que en dominios de imagen. El Synthetic Data Vault (SDV) introdujo un sistema para construir modelos generativos de bases de datos relacionales y para muestrear datos sintéticos de forma automatizada [16]. Posteriormente, CTGAN adaptó los modelos adversarios condicionales a datos tabulares y propuso estrategias para manejar variables continuas multimodales y categorías desbalanceadas [17].

Aunque los modelos profundos son atractivos, los entornos biomédicos con recursos limitados también requieren métodos transparentes, auditables y reproducibles. Por esa razón, este estudio utiliza una cópula gaussiana condicionada por clase como generador base. Esta familia de modelos separa la preservación de los marginales de la dependencia multivariada, permite inspeccionar la matriz de correlación aprendida y facilita la trazabilidad de las semillas, las particiones y los parámetros. En una implementación institucional, el generador puede reemplazarse por Red Antagonista Generativa Tabular Condicional (CTGAN), Autocodificador variacional tabular (TVAE), modelos bayesianos o mecanismos de privacidad diferencial, siempre que se mantenga la misma etapa de evaluación.

No obstante, la elección del generador no es neutral. En estudios comparativos con datos de pacientes, el desempeño relativo de distintos generadores varió según la métrica de utilidad o de privacidad considerada [18]; asimismo, se ha mostrado que la protección que ofrece un mismo método difiere entre registros, es particularmente débil para los valores atípicos y resulta difícil de anticipar sin una evaluación empírica [19]. Las revisiones sistemáticas sobre datos tabulares de salud documentan, además, una amplia diversidad de generadores y de criterios de evaluación, sin consenso

sobre el método más adecuado para cada caso [20]. Por estas razones, en este estudio la cópula gaussiana se adopta como línea base interpretable para demostrar el flujo de evaluación, no como generador preferente.

C. *Evaluación de utilidad y privacidad en salud*

La evaluación de datos sintéticos de salud suele organizarse en tres dimensiones: semejanza estadística, utilidad para tareas analíticas y privacidad [21]. La semejanza verifica si el conjunto sintético reproduce las distribuciones marginales, las correlaciones y las estructuras multivariadas. La utilidad mide si los datos sintéticos permiten obtener conclusiones o modelos comparables a los obtenidos con datos reales. La privacidad cuantifica si un conjunto sintético almacena registros o facilita inferencias sobre individuos. Estudios de validación de métricas han mostrado que la selección de indicadores debe alinearse con el caso de uso analítico, ya que una métrica global de similitud puede no reflejar el desempeño en una tarea biomédica específica [22].

Bajo esta lógica, el artículo adopta una evaluación mínima pero integrada: fidelidad marginal mediante el complemento de Kolmogorov-Smirnov, preservación de las correlaciones de Spearman, desempeño "train-synthetic-test-real" (TSTR), distancia al registro real más cercano (DCR), razón de vecino más cercano (NNDR), búsqueda de copias exactas y ataque de membresía basado en distancia. La selección busca equilibrar la interpretabilidad, el bajo costo computacional y la aplicabilidad en instituciones con infraestructura limitada.

III. Metodología propuesta

La metodología se diseñó como un ciclo de seis etapas, ilustrado en la Figura 1. La lógica central es que el dato sintético sólo debe considerarse apto para uso secundario cuando se demuestre que conserva suficiente utilidad para el caso científico definido y que no presenta señales relevantes de copia, cercanía excesiva o vulnerabilidad de membresía. El proceso no pretende reemplazar la evaluación ética, los contratos de uso ni los controles de seguridad; los complementa con evidencia cuantitativa, como se muestra en la Tabla I.

TABLA I
COMPONENTES OPERATIVOS DE LA METODOLOGÍA PROPUESTA

Etapa	Propósito	Salida verificable
Gobernanza y caso de uso	Define la finalidad, los responsables, las variables sensibles, el nivel de acceso y los criterios de aceptación.	Ficha de caso de uso y matriz de riesgos.
Preparación del dato real	Curación, codificación, tratamiento de faltantes, partición de entrenamiento/retención y eliminación de identificadores directos.	Conjunto real particionado y diccionario de variables.
Generación sintética	Aprende una distribución generativa y produce registros artificiales controlados por semilla y por la versión del modelo.	Conjunto sintético, parámetros y bitácora reproducibles.
Evaluación de utilidad	Mide la fidelidad marginal, la estructura multivariada y el desempeño TSTR en relación con la retención real.	Reporte de utilidad con umbrales de aceptación.
Evaluación de privacidad	Mide copia exacta, cercanía al registro real, NNDR y ataque de membresía.	Reporte de privacidad y clasificación de riesgo.
Decisión de liberación	Aprueba, restringe, regenera o rechaza el conjunto según su utilidad, riesgo y finalidad.	Acta de decisión y condiciones de uso.



Fig. 1 Flujo metodológico propuesto para la generación y evaluación de datos sintéticos biomédicos.

A. *Generación sintética condicionada*

Sea X una matriz de variables biomédicas tabulares y sea Y una variable objetivo clínica o epidemiológica. La partición inicial separa el conjunto de entrenamiento Dtrain y el conjunto de retención Dholdout. El generador se ajusta únicamente con Dtrain para evitar la fuga de información desde la partición utilizada para la evaluación. En la prueba de concepto se usó una cópula gaussiana por clase: para cada clase se transformaron las variables continuas a puntuaciones normales empíricas, se estimó una matriz de correlación regularizada y se muestrearon vectores multivariados normales que luego se devolvieron a la escala original mediante cuantiles empíricos.

El procedimiento preserva la distribución de clases de Dtrain de manera aproximada y genera n registros sintéticos, donde n coincide con el tamaño del conjunto de entrenamiento real. Esta decisión facilita comparar el desempeño predictivo en el escenario TSTR sin introducir ventajas artificiales derivadas del aumento del tamaño de la muestra. En contextos reales, n puede ajustarse según la finalidad, siempre que se reporte el cociente entre registros sintéticos y reales y se evalúe la sensibilidad de las métricas ante diferentes tamaños sintéticos.

B. *Métricas de utilidad*

La utilidad se midió en tres niveles. El primero fue la fidelidad marginal, calculada como el promedio del complemento de la estadística de Kolmogorov-Smirnov entre cada variable real de entrenamiento y su equivalente sintético. El segundo fue la similitud de correlaciones, calculada a partir del error absoluto medio entre las matrices de correlación de Spearman. El tercero fue la utilidad predictiva de TSTR: se entrenaron modelos con datos sintéticos y se evaluaron en un

conjunto de validación real. Como referencia superior, se calculó el train-real-test-real (TRTR). La pérdida de utilidad se definió como en la fórmula (1):

$$Delta\ AUC = AUC(TRTR) - AUC(TSTR). \quad (1)$$

Un delta AUC cercano a cero indica que el conjunto sintético conserva información predictiva relevante para la tarea definida. Sin embargo, una buena utilidad de TSTR no implica por sí misma la preservación de la interpretabilidad clínica, la calibración, la equidad por subgrupos ni la validez causal; por ello, la metodología debe ampliarse cuando el caso de uso involucre inferencia epidemiológica, decisiones clínicas o la asignación de recursos.

C. *Métricas de privacidad*

La privacidad se evaluó mediante métricas empíricas orientadas a detectar la memorización y la exposición individuales. Primero, se calculó la distancia euclidiana estandarizada desde cada registro sintético hasta su vecino más cercano en Dtrain y en Dholdout. Si los registros sintéticos se encuentran sistemáticamente mucho más cerca de Dtrain que de Dholdout, ello indica sobreajuste o copia. Segundo, se calculó la razón NNDR entre la distancia al vecino más cercano y la distancia al segundo vecino más cercano; valores muy bajos pueden señalar registros sintéticos excesivamente próximos a un individuo real específico. La Tabla II resume estas métricas y su interpretación operativa.

Tercero, se implementó un ataque de membresía basado en la distancia. Para cada registro candidato real, se calculó la distancia al registro sintético más cercano y se utilizó dicha distancia como puntuación de ataque para distinguir a los miembros de Dtrain de los no miembros de Dholdout. El área bajo la curva (AUC) del ataque se interpretó como un diagnóstico: un valor cercano a 0.5 sugiere un bajo poder de inferencia; valores claramente superiores a 0.5 requieren restringir o regenerar el conjunto de datos. Esta prueba no constituye una garantía formal de privacidad, pero sí un control práctico en contextos institucionales.

TABLA II
MÉTRICAS DE UTILIDAD Y PRIVACIDAD

Dimensión	Métricas	Interpretación operativa
Fidelidad	Complemento KS promedio	Valores cercanos a 1 indican una buena preservación de las distribuciones marginales.
Fidelidad	Similitud de correlación de Spearman	Valores cercanos a 1 indican la preservación de las dependencias monotónicas.
Utilidad	Delta AUC TRTR-TSTR	Pérdidas pequeñas sugieren que el dato sintético conserva una señal predictiva.
Utilidad	F1 y balanced accuracy TSTR.	Evalúan el desempeño en clases desequilibradas y los errores clínicamente relevantes.
Privacidad	DCR entrenamiento versus retención	Distancias excesivamente pequeñas al entrenamiento sugieren sobreajuste.
Privacidad	NNDR	Razones muy bajas indican una proximidad específica.
Privacidad	AUC de ataque de membresía	Cercano a 0.5 implica un bajo poder de inferencia; mayor riesgo si aumenta.
Privacidad	Copias exactas	Debe ser cero antes de cualquier liberación externa.

D. *Disponibilidad de datos y consideraciones éticas*

Este estudio utilizó exclusivamente un conjunto de datos público disponible en el UCI Machine Learning Repository. No se recolectaron datos primarios, no se accedió a expedientes clínicos identificables ni se procesó información personal sensible de participantes humanos. En aplicaciones con datos institucionales reales, la metodología debe someterse a aprobación ética, revisión legal y autorización del custodio de datos antes de cualquier síntesis o liberación.

IV. ESTUDIO DE PRUEBA DE CONCEPTO

La metodología se implementó con el conjunto de datos Wisconsin Diagnostic Breast Cancer (WDBC), disponible en el UCI Machine Learning Repository. El conjunto contiene 569 registros, una variable diagnóstica binaria y 30 variables reales calculadas a partir de características de núcleos celulares obtenidas a partir de imágenes digitalizadas de aspirado con aguja fina de masa mamaria [23]. Para evitar el uso de identificadores, se excluyó cualquier variable de identificación y se trabajó únicamente con características numéricas y la etiqueta diagnóstica. Aunque el conjunto es público, se eligió por su naturaleza biomédica, su tamaño moderado y su estructura tabular, que lo convierten en un caso útil para demostrar el flujo metodológico sin acceder a expedientes clínicos sensibles.

Esta elección tiene, sin embargo, una consecuencia que conviene explicitar: el WDBC es pequeño, carece de valores faltantes y está compuesto únicamente por variables continuas, de modo que plantea al generador un escenario favorable. No contiene variables categóricas de alta cardinalidad, registros longitudinales, eventos raros ni los cuasi-identificadores demográficos que concentran el riesgo de reidentificación en los expedientes reales [4], [5]. Por ello, los resultados deben leerse como verificación del funcionamiento del flujo de evaluación en condiciones controladas y no como estimación del desempeño esperable con datos institucionales sujetos a restricciones de privacidad.

Los datos se dividieron de forma estratificada en 70 % para entrenamiento real y 30 % para retención real. La partición de entrenamiento incluyó 398 registros y la de retención, 171. El generador sintético produjo 398 registros artificiales, lo que igualó el tamaño del conjunto de entrenamiento. La proporción de casos benignos fue de 0.628 en el conjunto de entrenamiento real y de 0.641 en el conjunto sintético, con una

diferencia absoluta pequeña. La semilla de reproducibilidad fue 2026. Los modelos de utilidad fueron regresión logística y bosque aleatorio; ambos se evaluaron en la misma partición de retención real para evitar sesgos de optimismo.

El experimento fue diseñado como validación de la metodología, no como un estudio clínico de desempeño diagnóstico. Por ello, las métricas no deben interpretarse como evidencia de la superioridad de un modelo para el diagnóstico del cáncer de mama, sino como demostración de cómo puede medirse la conservación de la señal estadística y predictiva al reemplazar el entrenamiento real por uno sintético.

V. Resultados

La Tabla III resume los resultados principales. La fidelidad marginal fue alta, con un complemento KS promedio de 0.954 y un valor mínimo de 0.937. La similitud de correlación de Spearman fue de 0.985, lo que sugiere la preservación de la estructura multivariada global. En utilidad predictiva, la regresión logística entrenada con datos reales obtuvo un AUC de 0.992, mientras que la entrenada con datos sintéticos obtuvo un AUC de 0.990 en la retención real, con una pérdida de 0.002. El bosque aleatorio mostró una pérdida de AUC de 0.004. Estos resultados indican que el conjunto sintético retuvo la principal señal discriminativa de la tarea definida. La Figura 2 presenta un resumen normalizado de estas métricas.

TABLA III
Resultados de utilidad y privacidad en la prueba de concepto

Bloque	Indicador	Resultado
Datos	Registros reales totales / entrenamiento / retención / sintéticos	569 / 398 / 171 / 398
Datos	Proporción de clase benigna en entrenamiento real / sintético	0.628 / 0.641
Fidelidad	Complemento KS promedio / mínimo	0.954 / 0.937
Fidelidad	Similitud de correlaciones de Spearman	0.985
Utilidad	Regresión logística: AUC TRTR / AUC TSTR / Delta AUC	0.992 / 0.990 / 0.002
Utilidad	Bosque aleatorio: AUC TRTR / AUC TSTR / Delta AUC	0.984 / 0.980 / 0.004
Utilidad	F1 TSTR regresión logística / bosque aleatorio	0.958 / 0.938
Privacidad	Mediana DCR sintético-entrenamiento / sintético-retención	2.288 / 2.506
Privacidad	Relación DCR entrenamiento-retención	0.913
Privacidad	Mediana NNDR	0.932
Privacidad	AUC ataque de membresía basado en distancia	0.512
Privacidad	Copias exactas detectadas	0.000

En la privacidad, no se identificaron copias exactas de registros reales. La mediana de la distancia desde registros sintéticos hasta entrenamiento real fue de 2.288, mientras que la distancia hasta retención real fue de 2.506. La relación DCR entrenamiento-retención fue de 0.913, lo que indica una mayor cercanía relativa al conjunto usado para entrenar el generador, pero no evidencia de colapso hacia copias exactas. La mediana de NNDR fue de 0.932, lo cual es consistente con la ausencia de vecinos únicos excesivamente dominantes. Finalmente, el ataque de membresía basado en distancia obtuvo un AUC de 0.512, cercano a la clasificación aleatoria. La Figura 3 muestra que las distribuciones de distancia de entrenamiento y retención se superponen ampliamente, aunque con una ligera tendencia a la cercanía al entrenamiento, esperable en un modelo ajustado sobre esa partición.

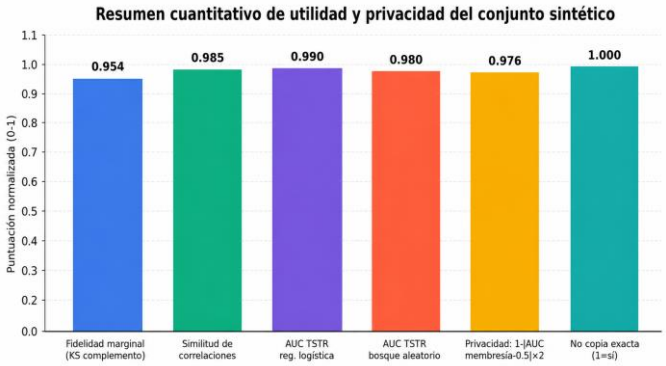


Fig. 2 Resumen normalizado de métricas de utilidad y privacidad del conjunto sintético.

En conjunto, la prueba de concepto alcanza un equilibrio favorable entre utilidad y privacidad empírica en un escenario de desarrollo metodológico. Parte de este resultado obedece a la naturaleza del conjunto: variables continuas, completas y de distribución relativamente regular se ajustan bien a una cópula gaussiana, de modo que la fidelidad observada no puede atribuirse únicamente al generador ni suponerse transferible a datos de otra naturaleza. No obstante, el resultado no debe extrapolarse automáticamente a expedientes electrónicos, datos genómicos, series temporales, imágenes médicas, datos raros ni cohortes pequeñas. En esos dominios, la probabilidad de singularidad individual y de memorización puede ser mayor, por lo que se requieren métricas adicionales, revisión experta y, en algunos casos, entrenamiento con privacidad diferencial.

VI. Discusión

Los resultados respaldan la viabilidad de una metodología de datos sintéticos evaluables para una investigación biomédica segura. La principal implicación científica es que la discusión no debe centrarse en si los datos sintéticos son útiles o privados en abstracto, sino en si un conjunto sintético específico, producido por un generador específico y destinado a un caso de uso específico, supera umbrales predefinidos de utilidad y privacidad. Esta postura evita dos errores frecuentes: liberar datos sintéticos sin evaluación porque parecen

artificiales, o descartar la síntesis de datos por no ofrecer anonimato absoluto.

Desde una perspectiva tecnológica y de negocios, la metodología puede convertirse en un servicio institucional o en una plataforma interna de "data sharing seguro". Un hospital universitario, una secretaría de salud o una red de investigación podría ofrecer a terceros un catálogo de conjuntos sintéticos, cada uno acompañado de un diccionario de variables, un reporte de calidad, una tarjeta de riesgos, una versión del generador, una semilla y las condiciones de uso. Para emprendimientos de salud digital, este enfoque permitiría probar software, validar canalizaciones de interoperabilidad, entrenar equipos de ciencia de datos y desarrollar modelos preliminares sin requerir acceso inicial a expedientes reales. Para investigadores, reduciría los tiempos de negociación de datos y facilitaría la reproducibilidad antes de pasar a entornos cerrados con datos reales.

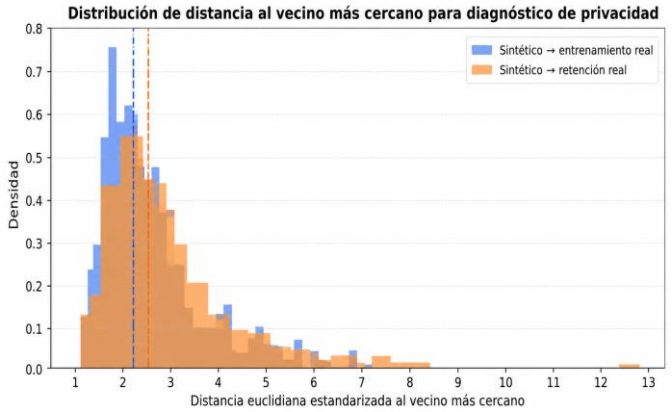


Fig. 3 Distribución de la distancia al vecino más cercano desde registros sintéticos hasta entrenamiento y retención reales.

La propuesta también tiene valor en contextos latinoamericanos, donde la fragmentación de los sistemas de información, la baja interoperabilidad y la ausencia de infraestructuras robustas de cómputo seguro dificultan los proyectos colaborativos. Un marco de síntesis evaluable puede operar como tecnología intermedia: no exige liberar datos originales ni construir de inmediato un repositorio clínico nacional, pero permite generar evidencia inicial, entrenar talento y diseñar productos analíticos con menor exposición de datos personales. La adopción debe ir acompañada de políticas de gobernanza, revisión legal, criterios de acceso escalonado y capacitación de los comités éticos para la evaluación de riesgos algorítmicos.

Otro aspecto relevante es el riesgo de que los datos sintéticos generen una falsa percepción de anonimato absoluto. En entornos institucionales, puede asumirse erróneamente que un conjunto artificial deja de ser sensible automáticamente, lo que reduce los controles de gobernanza, auditoría o evaluación ética. Sin embargo, estudios recientes han demostrado que ciertos modelos generativos pueden memorizar patrones individuales o facilitar inferencias indirectas cuando se dispone de fuentes de información auxiliares [5], [14], [24]. Por ello, la síntesis de datos debe entenderse como una estrategia de reducción de riesgos y no como un sustituto del cumplimiento normativo, de la supervisión organizacional o de la evaluación continua de la privacidad.

La prueba de concepto presentó limitaciones. Primero, utilizó un conjunto público relativamente pequeño, limpio y numérico; los expedientes electrónicos reales presentan faltantes de información, variables temporales, eventos raros, múltiples tablas y sesgos institucionales; por ello, persiste una distancia considerable entre el caso evaluado y los escenarios de privacidad restringida que motivan la propuesta. Segundo, el generador de cópula gaussiana es transparente, pero puede subrepresentar dependencias no lineales complejas. Además, no se comparó con generadores alternativos, como CTGAN, TVAE, redes bayesianas o generadores con privacidad diferencial, de modo que los resultados no permiten afirmar que la cópula sea preferible a otras opciones. Tercero, las métricas de privacidad implementadas son diagnósticas y no constituyen garantías formales. Cuarto, el ataque de membresía basado en la distancia es una aproximación simple; adversarios con conocimiento auxiliar podrían emplear estrategias más fuertes. Quinto, la utilidad clínica no fue evaluada por especialistas ni mediante desenlaces prospectivos. Sexto, el análisis empleó una única partición y una única semilla, por lo que no se estimó la variabilidad de las métricas entre repeticiones.

Asimismo, la preservación de la utilidad global no garantiza una representación adecuada de subgrupos clínicos minoritarios o de poblaciones históricamente subrepresentadas. Un conjunto sintético puede mantener métricas agregadas aceptables y, al mismo tiempo, distorsionar patrones relevantes para enfermedades raras, grupos étnicos específicos o poblaciones con baja frecuencia estadística. Este problema es especialmente importante en salud, donde pequeñas desviaciones pueden afectar los modelos predictivos, la priorización clínica o la evaluación epidemiológica [25], [26]. En consecuencia, futuras implementaciones deberían incorporar auditorías de equidad y evaluaciones estratificadas por subgrupos antes de utilizar datos sintéticos para el entrenamiento o la validación de sistemas clínicos basados en inteligencia artificial.

Estas limitaciones no invalidan el enfoque; más bien definen una agenda de maduración. La siguiente etapa de validación debería aplicar el mismo protocolo, con particiones y umbrales fijados de antemano, a conjuntos de complejidad creciente: datos tabulares mixtos con variables categóricas y valores faltantes, cohortes pequeñas con eventos raros y, finalmente, expedientes electrónicos longitudinales y multitabla bajo acceso controlado. En cada nivel, la cópula gaussiana debería contrastarse con CTGAN, TVAE, modelos bayesianos y al menos un generador con privacidad diferencial, reportando la media y la dispersión de las métricas en varias semillas. Marcos abiertos de evaluación comparativa, como Synthcity [27], permiten realizar este

contraste sin modificar la etapa de evaluación propuesta. En aplicaciones reales se recomienda, incorporar privacidad diferencial cuando el riesgo sea alto, auditar subgrupos minoritarios, medir la calibración y la equidad, evaluar la utilidad específica del estudio y documentar los umbrales de aceptación antes de generar el dato sintético. Además, los conjuntos sintéticos deben clasificarse por niveles de acceso: uso interno, "sandbox" controlado, colaboración bajo convenio o liberación pública. El nivel de acceso debe depender tanto de las métricas como de la sensibilidad del dominio.

VII. Implicaciones para implementación institucional

La implementación organizacional requiere convertir la metodología en un proceso bien gobernado. Se propone crear un comité técnico-jurídico que revise las finalidades, determine variables no sintetizables, defina umbrales mínimos y autorice niveles de acceso [28], [21]. Cada liberación debe incluir una ficha técnica con el origen del dato, transformaciones, modelo generativo, métricas, fecha, responsable, restricciones de uso y advertencias sobre la no utilización clínica directa [21]. Este enfoque reduce la ambigüedad y permite auditar las decisiones ex post [28].

En términos de arquitectura, la generación debe ejecutarse en un entorno seguro donde permanezca el dato real [29]. Sólo el conjunto sintético y su reporte deberían salir del entorno, y únicamente después de superar los controles de utilidad y de privacidad [10]. Para instituciones con recursos limitados, el prototipo puede operar con Python, bibliotecas abiertas y almacenamiento local cifrado; a medida que aumenta su madurez, puede migrar a contenedores reproducibles, control de versiones de los datos, orquestación de pipelines y paneles de monitoreo de riesgo [28], [29]. La inversión inicial es menor que la necesaria para desplegar un data lake clínico completo, pero genera beneficios tempranos en innovación y colaboración [29].

VIII. Conclusiones

Los datos sintéticos ofrecen una alternativa práctica para reducir las barreras de acceso a los datos biomédicos, siempre que su generación vaya acompañada de una evaluación clara de su utilidad y de su privacidad. En este trabajo se presentó una metodología reproducible que combina gobernanza de datos, síntesis tabular y métricas orientadas tanto a preservar el valor analítico de la información como a identificar posibles riesgos para la privacidad. La prueba de concepto realizada con el conjunto de datos WDBC mostró alta fidelidad estadística, mínima pérdida predictiva y bajo poder de inferencia de membresía, lo que confirma la operatividad del flujo en un caso controlado. Puesto que dicho caso es pequeño, completo y exclusivamente numérico, la extensión del enfoque a escenarios de investigación biomédica con restricciones reales de acceso a los datos deberá demostrarse con conjuntos más heterogéneos y mediante la comparación con generadores alternativos.

La conclusión principal es que los datos sintéticos no deben considerarse automáticamente seguros solo por ser artificiales. Su uso responsable requiere validación cuantitativa, trazabilidad y supervisión institucional antes de cualquier liberación o reutilización. Bajo este enfoque, la síntesis evaluable puede facilitar la colaboración científica, la formación, la innovación y el desarrollo temprano de soluciones basadas en inteligencia artificial sin exponer datos reales sensibles. Como línea futura, será necesario extender la metodología a datos longitudinales, expedientes electrónicos multitabla, imágenes médicas, datos genómicos y modelos entrenados con garantías formales de privacidad diferencial, así como realizar una evaluación comparativa sistemática entre generadores bajo un protocolo común.

Agradecimientos

Agradecemos el apoyo recibido del Instituto de Investigación en Ciencias Médicas y Derecho a la Salud (ICIMEDES) de la Facultad de Ciencias Médicas (FCM) y de la Dirección de Investigaciones Científicas, Humanísticas y Tecnológicas (DICIHT), todas ellas de la Universidad Nacional Autónoma de Honduras (UNAH).

Referencias

[1] M. D. Wilkinson *et al*., "The FAIR Guiding Principles for scientific data management and stewardship", *Sci. Data*, vol. 3, Art. no. 160018, Mar. 2016, doi: 10.1038/sdata.2016.18

[2] B. Lo, "Sharing clinical trial data: Maximizing benefits, minimizing risk", *JAMA*, vol. 313, no. 8, pp. 793–794, Feb. 2015, doi: 10.1001/jama.2015.292

[3] L. Sweeney, "k-anonymity: A model for protecting privacy", *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, Oct. 2002, doi: 10.1142/S0218488502001648

[4] K. El Emam, E. Jonker, L. Arbuckle, and B. Malin, "A systematic review of re-identification attacks on health data", *PLoS ONE*, vol. 6, no. 12, Art. no. e28071, Dec. 2011, doi: 10.1371/journal.pone.0028071

[5] L. Rocher, J. M. Hendrickx, and Y.-A. de Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models", *Nat. Commun.*, vol. 10, Art. no. 3069, Jul. 2019, doi: 10.1038/s41467-019-10933-3

[6] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich, "Identifying personal genomes by surname inference", *Science*, vol. 339, no. 6117, pp. 321–324, Jan. 2013, doi: 10.1126/science.1229566

[7] Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. S. Pentland, "Unique in the shopping mall: On the reidentifiability of credit card metadata", *Science*, vol. 347, no. 6221, pp. 536–539, Jan. 2015, doi: 10.1126/science.1256297

[8] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement", *Ann. Intern. Med.*, vol. 162, no. 1, pp. 55–63, Jan. 2015, doi: 10.7326/M14-0697

[9] A. Gonzales, G. Guruswamy, and S. R. Smith, "Synthetic data in health care: A narrative review", *PLOS Digit. Health*, vol. 2, no. 1, Art. no. e0000082, Jan. 2023, doi: 10.1371/journal.pdig.0000082.

[10] U.S. Department of Health and Human Services, "Guidance regarding methods for de-identification of protected health information in

accordance with the HIPAA Privacy Rule", Feb. 3, 2025. [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html. Accessed: May 10, 2026.
[11] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016", *Off. J. Eur. Union*, vol. L119, pp. 1–88, May 4, 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj. Accessed: May 10, 2026.
[12] S. L. Garfinkel, *De-Identification of Personal Information*, Natl. Inst. Stand. Technol., Gaithersburg, MD, USA, NISTIR 8053, Oct. 2015, doi: 10.6028/NIST.IR.8053.
[13] C. Dwork, "Differential privacy", in *Proc. 33rd Int. Colloq. Automata Lang. Program. (ICALP)*, Venice, Italy, 2006, pp. 1–12, doi: 10.1007/11787006_1.
[14] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models", in *Proc. IEEE Symp. Secur. Privacy (SP)*, San Jose, CA, USA, 2017, pp. 3–18, doi: 10.1109/SP.2017.41.
[15] I. J. Goodfellow et al., "Generative adversarial nets", in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, Montreal, QC, Canada, 2014, pp. 2672–2680.
[16] N. Patki, R. Wedge, and K. Veeramachaneni, "The Synthetic Data Vault", in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Montreal, QC, Canada, 2016, pp. 399–410, doi: 10.1109/DSAA.2016.49.
[17] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN", in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Vancouver, BC, Canada, 2019, pp. 7335–7345.
[18] A. Goncalves, P. Ray, B. Soper, J. Stevens, L. Coyle, and A. P. Sales, "Generation and evaluation of synthetic patient data," *BMC Med. Res. Methodol.,* vol. 20, Art. no. 108, May 2020, doi: 10.1186/s12874-020-00977-1.
[19] T. Stadler, B. Oprisanu, and C. Troncoso, "Synthetic data – Anonymisation groundhog day," in *Proc. 31st USENIX Secur. Symp*., Boston, MA, USA, 2022, pp. 1451–1468.
[20] M. Hernandez, G. Epelde, A. Alberdi, R. Cilla, and D. Rankin, "Synthetic data generation for tabular health records: A systematic review," *Neurocomputing,* vol. 493, pp. 28–45, Jul. 2022, doi: 10.1016/j.neucom.2022.04.053.
[21] M. Hernandez, G. Epelde, A. Alberdi, R. Cilla, and D. Rankin, "Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimensions", *Methods Inf. Med., vol.* 62, no. S 01, pp. e19–e38, Jan. 2023, doi: 10.1055/s-0042-1760247.
[22] K. El Emam, L. Mosquera, X. Fang, and A. El-Hussuna, "Utility metrics for evaluating synthetic health data generation methods: Validation study", *JMIR Med. Inform.*, vol. 10, no. 4, Art. no. e35734, Apr. 2022, doi: 10.2196/35734.
[23] W. Wolberg, O. Mangasarian, N. Street, and W. Street, *Breast Cancer Wisconsin (Diagnostic)*. Irvine, CA, USA: UCI Machine Learning Repository, 1993, doi: 10.24432/C5DW2B.
[24] N. Papernot, M. Abadi, Ú. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," in Proc. Int. Conf. Learn. Representations (ICLR): Toulon, France, 2017. [Online]. Available https://arxiv.org/abs/1610.05755
[25] I. Chen, F. D. Johansson, and D. Sontag, "Why is my classifier discriminatory?", in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 31, Montreal, QC, Canada, 2018. [Online[Available: https://arxiv.org/abs/1805.12002
[26] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations", Science, vol. 366, no. 6464, pp. 447–453, Oct. 2019, doi: 10.1126/science.aax2342.
[27] Z. Qian, B.-C. Cebere, and M. van der Schaar, "Synthcity: Facilitating innovative use cases of synthetic data in different data modalities," 2023, arXiv:2301.07573.
[28] T. Gebru *et al*., "Datasheets for datasets," *Commun. ACM*, vol. 64, no. 12, pp. 86–92, Dec. 2021, doi: 10.1145/3458723
[29] P. R. Burton *et al*., "Data Safe Havens in health research and healthcare," *Bioinformatics*, vol. 31, no. 20, pp. 3241–3248, Oct. 2015, doi: 10.1093/bioinformatics/btv279.