Diseño de un modelo para generar datos sintéticos en investigación médica.

Alfonso Esteban Lasso[1, 2, a], Cristina Martínez Toledo[1], Sergio Perosanz Amarillo[3]

1. Inari Biotech. **2.** Dpto. Biomedicina y Biotecnología, Universidad de Alcalá. **3.** Depto. de Materiales y Producción Aeroespacial, ETSI Aeronáutica y del Espacio, Universidad Politécnica de Madrid.

a. inaribiotech@gmail.com

VIII Congreso de Señalización Celular, SECUAH 2023.
21 a 24 de marzo, 2023. Universidad de Alcalá. Alcalá de Henares, Madrid. España.

Palabras clave: Inteligencia artificial; datos sintéticos; ensayos clínicos; GANs; VAES

Resumen

La escasez de datos o una distribución irregular, con espacios poco poblados de datos, son dificultades frecuentes de la investigación biomédica. Esta escasez de información dificulta el trabajo estadístico, impidiendo la obtención de conclusiones válidas. Las enfermedades raras, que se definen como aquellas enfermedades cuya prevalencia es inferior a 5 por cada 10.000 personas, son un ejemplo paradigmático de escasez de datos debido a su baja prevalencia. Actualmente es posible solucionar parcialmente este problema mediante la utilización de datos sintéticos. Los datos sintéticos son datos fiables generados mediante inteligencia artificial que complementan a los datos reales cuando los conjuntos de datos reales carecen de calidad, volumen o variedad. En este estudio se compararon la calidad de los datos sintéticos de investigación médica obtenidos mediante distintos algoritmos generadores de datos sintéticos. Se utilizaron datos tabulares de ensayos clínicos oncológicos procedentes del banco de datos de Project Data Sphere (PDS) y del National Cancer Institute. Algunos de los indicadores usados para comparar los datos sintéticos con los reales y evaluar la calidad de los datos sintéticos generados fueron: precisión, recall y F1score. Según nuestro estudio, la puntuación de los algoritmos menos intensivos en computación era relativamente baja; 63% para GaussianCopula, y 68% para Fast-ML. Los algoritmos más sofisticados generaron una mejor puntuación promedio; 74% para CTGAN, 78% para CopulaGAN y 82% para TVAE. En conclusión, los algoritmos TVAE son los más idóneos para generar datos sintéticos de datos médicos y serian de utilidad, por ejemplo, para generar datos médicos de pacientes con enfermedades raras que aumenten la base de estudio. Además, se desarrolló un nuevo algoritmo sobre la base de TVAE para su uso en investigación médica y en ensayos clínicos. El algoritmo se probó para generar datos sintéticos con los datos del ensayo N0147 de PDS. Este trabajo demuestra que se obtienen las mismas conclusiones analizando los datos sintéticos que analizando los datos reales lo cual sugiere que los datos sintéticos podrían ser de utilidad en investigación médica. Inari agradece a la CAM la concesión de ayudas del programa Investigo.

Cita: Esteban Lasso, Alfonso; Martínez Toledo, Cristina; Perosanz Amarillo, Sergio (2023) Diseño de un modelo para generar datos sintéticos en investigación médica. Actas del VIII Congreso de Señalización Celular, SECUAH 2023. 21 a 24 de marzo, 2023. Universidad de Alcalá. Alcalá de Henares, Madrid. España. dianas 12 (1): e202303fp01. ISSN 1886-8746 (electronic) journal.dianas.e202303fp01 https://dianas.web.uah.es/journal/e202303fp01. URI http://hdl.handle.net/10017/15181

Introducción

La escasez de datos o una distribución irregular, con espacios poco poblados de datos, son dificultades frecuentes de la investigación biomédica. En ocasiones, esta escasez de información dificulta el trabajo estadístico, impidiendo la obtención de conclusiones válidas. Las enfermedades raras, que se definen como aquellas enfermedades cuya prevalencia es inferior a 5 por cada 10.000 personas, son un ejemplo paradigmático de escasez de datos, aunque existen múltiples situaciones en las que se produce esta ausencia de datos científicos. Actualmente es posible solucionar, parcialmente, este problema mediante la utilización de datos sintéticos. Los datos sintéticos son datos fiables generados mediante inteligencia artificial que complementan a los datos reales cuando los conjuntos de datos reales carecen de calidad, volumen o variedad. Estos datos se usan en aplicaciones como la minería de datos, el aprendizaje automático, la simulación de sistemas complejos y la protección de la privacidad. En biomedicina, los datos sintéticos se utilizan para anonimizar los datos reales de los pacientes de instituciones sanitarias y que puedan ser usados en investigación de forma ágil y eficaz a la vez que se cumple la legislación de protección de datos [1-3]. Los datos sintéticos suponen la mejor solución al reto de la protección de datos médicos ya que no hay información de identificación que pueda rastrearse hasta pacientes individuales. Además, los pacientes sintéticos anonimizados son útiles para estudios de reposicionamiento y estudios secundarios de fármacos [1, 4].

Existen varios algoritmos para producir datos sintéticos como VAE (Variational Autoencoders) o GANs (Generative Adversarial Network). La primera GAN fue desarrollada en 2014 por Goodfellow y colaboradores [5]. Una GAN consiste en dos redes neuronales que compiten entre sí en un juego minimax. Las redes, denominadas Generador y Discriminador, respectivamente, tratan de cumplir tareas diferentes: mientras que el Generador pretende crear nuevos puntos de datos que sean lo más parecidos posible a los originales, el objetivo del Discriminador es identificar correctamente esas falsificaciones creadas artificialmente. A lo largo del entrenamiento del modelo, ambas redes se vuelven más precisas en la resolución de sus respectivas tareas, hasta que alcanzan un equilibrio en el que los objetos generados ya no pueden distinguirse de los originales. La Figura 1 ilustra su arquitectura y aplicación para el diseño *in silico* de nuevas moléculas en farmacología clínica [6]. Aunque las GANs son más conocidas por crear objetos sintéticos basados en datos no estructurados, como imágenes, también muestran resultados prometedores en datos estructurados y tabulares. Las GANs bien entrenadas no sólo pueden aumentar el tamaño de la muestra de estudios pequeños, sino que también pueden mejorar la disponibilidad de datos de pacientes con valores extremos, facilitando, entre otros, los análisis exploratorios de subgrupos. Además, su arquitectura especial convierte a las GANs en una solución ideal para los problemas de privacidad de los datos: aunque el algoritmo Discriminador se entrena utilizando datos originales, el Generador no tiene acceso directo a esta información, mejorando su proceso generativo únicamente a través de la retroalimentación del Discriminador.

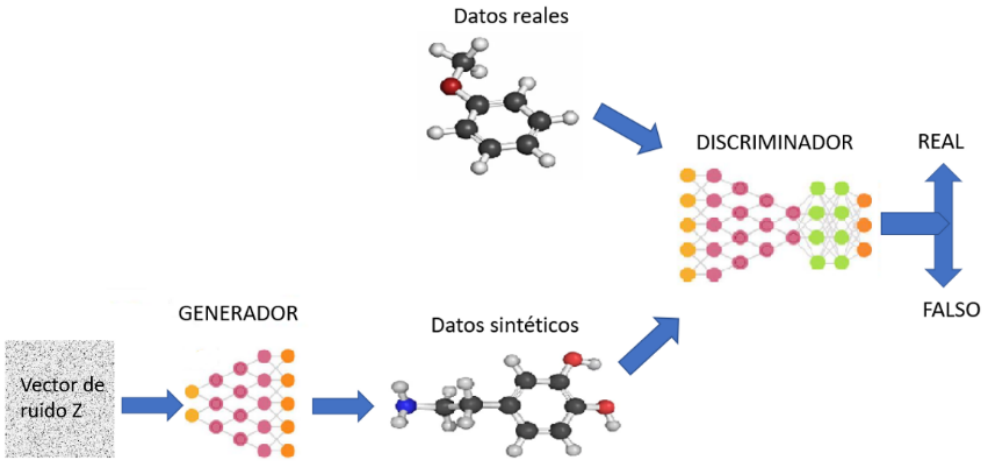


Figura 1: arquitectura del algoritmo GAN

Por otro lado, VAE es un algoritmo generador de datos, el cual busca estimar la función de densidad de probabilidad de los datos reales. Esto se logra produciendo dos vectores, uno con promedios (μ), y otro con las desviaciones estándar (σ) desde los datos reales. El algoritmo intenta aprender de las distribuciones de las variables latentes o inferidas, basándose en el promedio y la varianza por medio de compresiones y descompresiones de la información, utilizando un autocodificador o autoencoder.

A diferencia de los objetos no estructurados, los datos clínicos estructurados se basan en dependencias de datos muy específicas que deben ser tenidas en cuenta por los algoritmos generativos para tener éxito, una tarea que no se ha resuelto satisfactoriamente hasta la fecha [7]. En este estudio se compararon distintos algoritmos generadores de datos sintéticos determinando la calidad de los datos sintéticos generados a partir de datos tabulares de ensayos clínicos. Se desarrolló un software (sobre la base del algoritmo que proporcionó mejores resultados) y se probó para determinar su utilidad en el análisis de datos de ensayos clínicos.

Material y métodos

Para este estudio se utilizaron los datos del ensayo clínico oncológico N0147 del repositorio Project Data Sphere (PDS), tras obtener la aprobación del National Cancer Institute. Este ensayo ha sido objeto de tres estudios previos: el ensayo clínico original [8], un estudio retrospectivo secundario [9], y un análisis para determinar si se obtenían las mismas conclusiones del estudio secundario con datos sintéticos [4]. Se seleccionó este ensayo para determinar si con los datos sintéticos generados se llegaba a las mismas conclusiones que los estudios previos [4,9], replicando sus análisis e interpretando los datos de la misma manera.

Los algoritmos generadores de datos sintéticos analizados en este estudio fueron: GaussianCopula, Fast-ML, CTGAN, CopulaGAN y TVAE; seleccionando como base para nuestro análisis, en última instancia, el TVAE por sus altos porcentajes de fiabilidad (>82%). Para poder medir el desempeño de los modelos y evaluar la calidad de los datos generados, se usaron diferentes métricas de rendimiento: Accuracy, Recall, Precision o F1-Score.

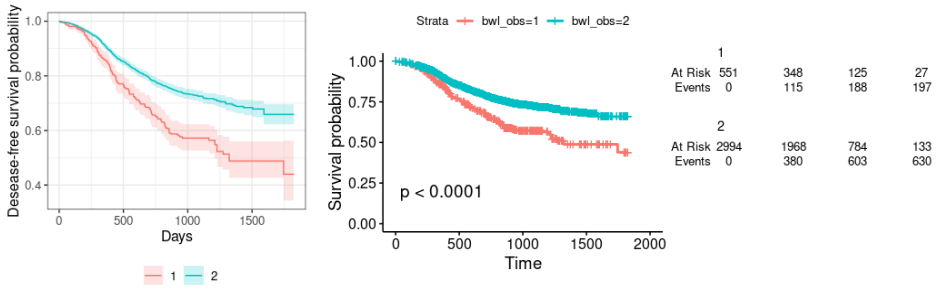
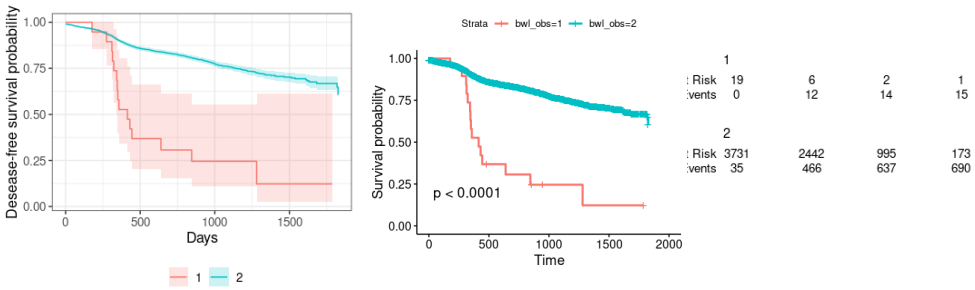


Figura 2.- Datos reales. Curvas de supervivencia libre de enfermedad, p-valor al análisis de supervivencia y tabla de riesgos correspondiente a este. 1=sin obstrucción. 2=con obstrucción.



Figuras 3.- Datos sintéticos. Curvas de supervivencia libre de enfermedad, p-valor al análisis de supervivencia y tabla de riesgos correspondiente a este. 1=sin obstrucción. 2=con obstrucción.

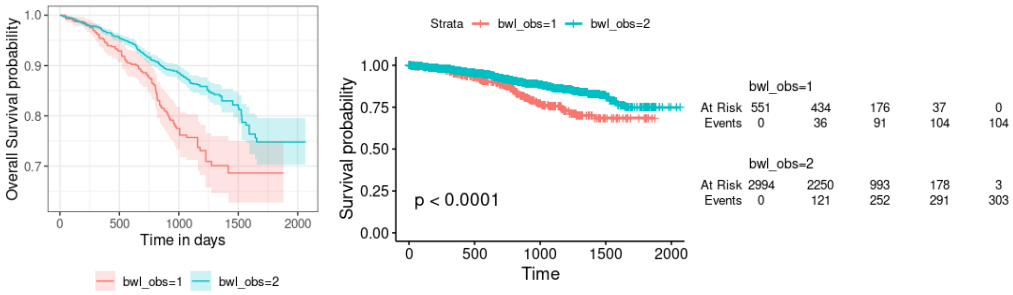


Figura 4.- Datos reales. Curvas de supervivencia total, p-valor al análisis de supervivencia y tabla de riesgos correspondiente a este. 1=sin obstrucción. 2=con obstrucción.

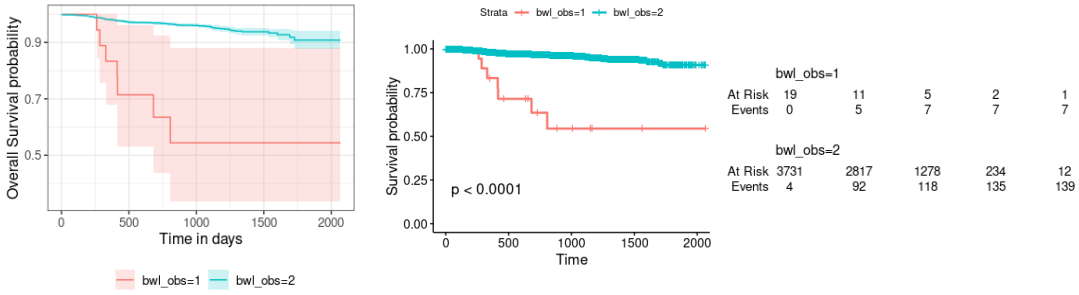


Figura 5.- Datos sintéticos. Curvas de supervivencia total, p-valor al análisis de supervivencia y tabla de riesgos correspondiente a este. 1=sin obstrucción. 2=con obstrucción.

Resultados y discusión

Según los resultados obtenidos sobre la calidad de los datos sintéticos, la puntuación de los algoritmos menos intensivos en computación era relativamente baja; 63% para GaussianCopula, y 68% para Fast-ML. Los algoritmos más sofisticados generaron una mejor puntuación promedio; 74% para CTGAN, 78% para CopulaGAN y 82% para TVAE. En conclusión, el algoritmo TVAE es el más idóneo, como base, para generar datos sintéticos a partir de los datos tabulares de ensayos clínicos. TVAE es una adaptación del algoritmo VAE a los datos tabulares, creado en 2019, que según sus desarrolladores es superior a la mayoría de los algoritmos en la generación de datos sintéticos [10]. Los distintos estudios publicados sobre la calidad de los datos generados, ni usan los mismos conjuntos de datos, ni las mismas métricas de calidad, por lo que es difícil la comparación de los resultados publicados. En nuestro caso, este estudio nos permitió seleccionar el algoritmo TVAE como punto de partida con el objetivo de mejorarlo para desarrollar un modelo generador de datos sintéticos para los datos tabulares de ensayos clínicos.

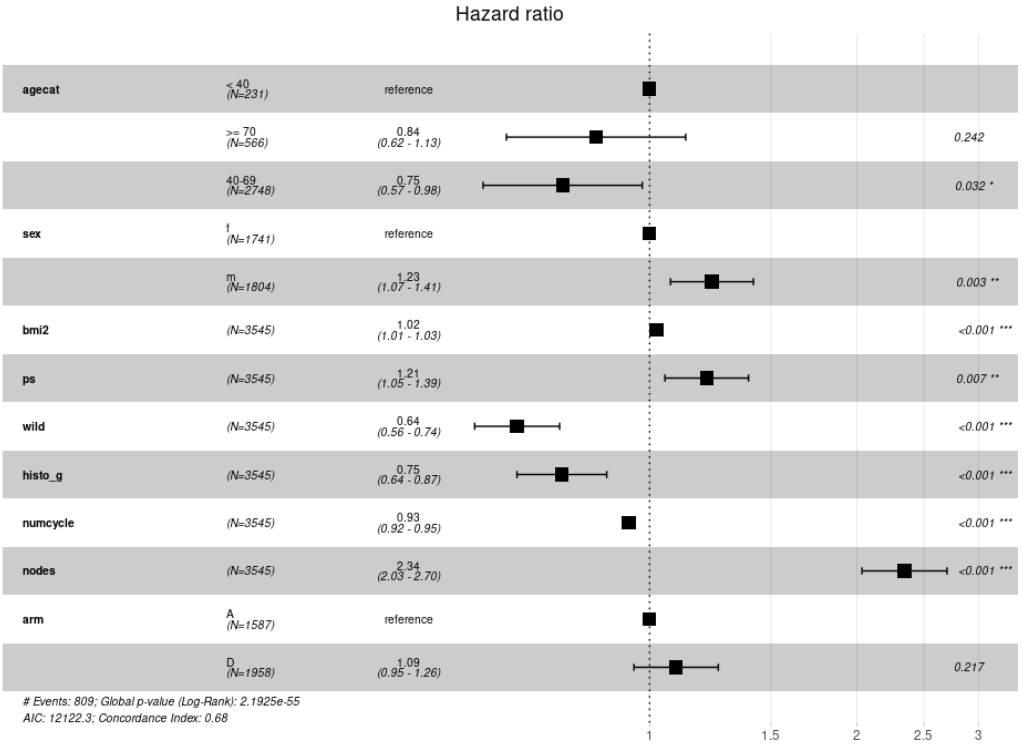


Figura 6.- Tabla de Hazard Ratio para datos reales, p-valor asociado a la regresión Cox e intervalos de confianza asociados a la supervivencia libre de enfermedad.

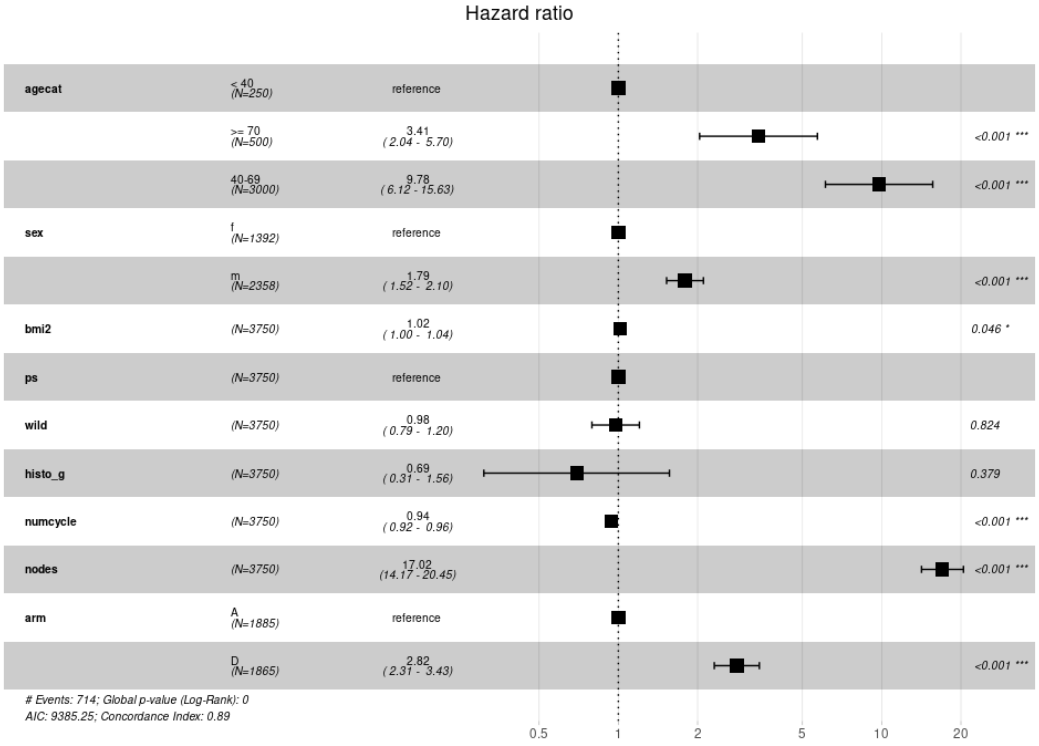


Figura 7.- Tabla de Hazard Ratio para datos sintéticos, p-valor asociado a la regresión Cox e intervalos de confianza asociados a la supervivencia libre de enfermedad.

El algoritmo TVAE inicial proporcionaba datos sintéticos con una valoración media del 82%. Después de mejorar y depurar el software se obtuvo una puntuación media del 87-90% con picos puntuales en algunas pruebas realizadas del 93,69%. Una vez alcanzado este nivel de calidad en los datos sintéticos, el siguiente objetivo fue evaluar su utilidad, para lo cual es fundamental no solo la calidad sino también la coherencia entre las distintas variables. Para ello comparamos los resultados y conclusiones del ensayo N0147 usando datos reales con los resultados que se obtiene usando datos sintéticos. El ensayo clínico original N0147, se realizó entre el 2004 y el 2009 con 2.527 pacientes con adenocarcinoma de colon en estadio III, que tras la resección quirúrgica recibían tratamiento quimioterápico FOLFOX con cetuximab y sin cetuximab (grupo control). La asignación de paciente a uno u otro grupo se realizó de forma aleatoria en un ensayo a doble ciego y se analizaba como criterio de valoración clínica (endpoint) la supervivencia libre de enfermedad [8]. El estudio retrospectivo secundario, se centraba en los 1.543 pacientes del grupo control al objeto de

determinar si la obstrucción intestinal en estos pacientes es un factor de mal pronóstico. Se analizaban, como endpoints, el tiempo hasta la recurrencia y la supervivencia general [9]. La principal conclusión del estudio secundario fue que la ausencia de obstrucción intestinal tiene un fuerte impacto en la supervivencia de los pacientes. Finalmente, Azizi y cols [4] estudiaron si con los datos sintéticos generados mediante los algoritmos del tipo arboles de decisión (conditional trees) llegaba a las mismas conclusiones del estudio secundario [9].

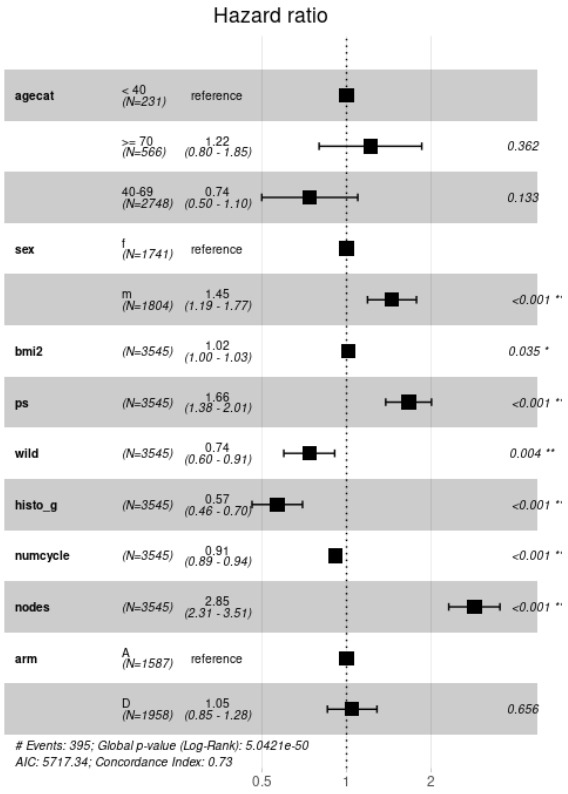


Figura 8.- Tabla de Hazard Ratio para datos reales, p-valor asociado a la regresión Cox e intervalos de confianza asociados a la supervivencia general.

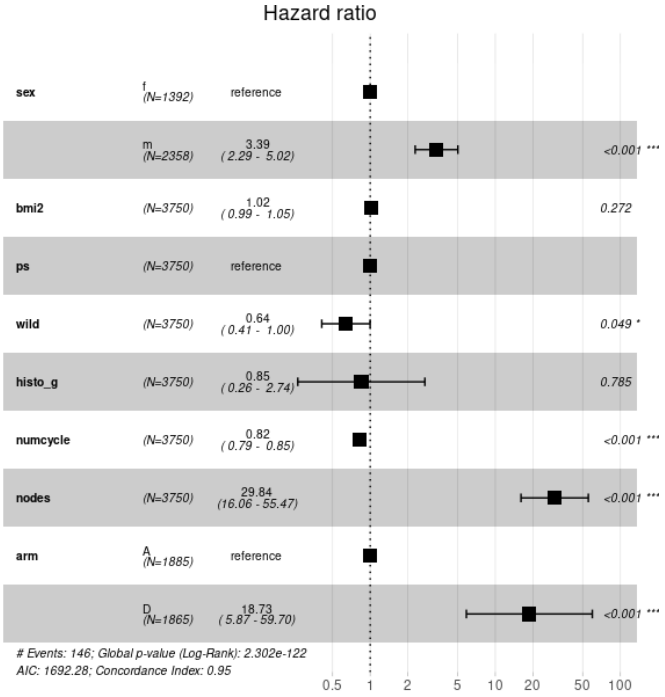


Figura 9.- Tabla de Hazard Ratio para datos sintéticos, p-valor asociado a la regresión Cox e intervalos de confianza asociados a la supervivencia general.

En nuestro caso, tras representar gráficamente los análisis de supervivencia, tanto para la supervivencia libre de enfermedad (Figuras 2 y 3) como para la supervivencia general (Figuras 4 y 5), hemos obtenido unos resultados que, aunque algo distintos a los obtenidos con los datos reales (debido a que software se

encuentra todavía en una fase temprana de desarrollo), llegan con un nivel de significancia igual a los reales y con tendencias que sugieren las mismas conclusiones.

Para el análisis multivariante, Azizi y cols. usaron el modelo de regresión Cox, que es la regresión comúnmente usada para analizar factores pronósticos y estimar ratios de supervivencia. Por ello este fue el análisis realizado en el presente trabajo. En el estudio de Azizi y cols. la coincidencia entre los datos reales y sintéticos para la variable de supervivencia general fue del 61% [datos reales: HR (Hazard Ratio) 1.56; 95% CI (Intervalo de confianza) 1.11 a 2.2; Datos sintéticos: HR 2.03; 95% CI 1.44 a 2.87]. Para la variable de supervivencia libre de enfermedad la coincidencia entre los datos reales y sintéticos fue del 86% [datos reales: HR 1.51; 95% CI: 1.18 a 1.95, Datos sintéticos: HR 1.63; 95% CI 1.26 a 2.1].

Los resultados obtenidos con los algoritmos TVAE fueron para la supervivencia general: [datos reales: HR (Hazard Ratio) 1.05; 95% CI (Intervalo de confianza) 0.85 a 1.28 e índice de concordancia 0,68 (Figura 6); Datos sintéticos: HR 18.73; 95% CI 5.87 a 59.70 e índice de concordancia 0,89 (Figura 7)] y para el tiempo libre de enfermedad: [datos reales: HR (Hazard Ratio) 1.09; 95% CI (Intervalo de confianza) 0.95 a 1.26 e índice de concordancia 0,73 (Figura 8); Datos sintéticos: HR 2,82; 95% CI 2.31 a 3.43 e índice de concordancia 0,95 (Figura 9)].

Todavía son escasos los estudios orientados a validar la fiabilidad de los datos sintéticos en investigación médica. En un estudio que comparaba los resultados obtenidos a partir de datos reales de 5 estudios clínicos con los resultados obtenidos a partir de los respectivos datos sintéticos generados mediante la aplicación comercial MDClone, se demostró que los resultados obtenidos con los datos sintéticos eran altamente predictivos de los resultados obtenidos con datos reales [2]. Similares conclusiones se han obtenido utilizando datos de ensayos clínicos y generando los datos sintéticos con algoritmos del tipo conditional trees [4] o GAN [1]. Nuestro estudio, aunque preliminar, viene a confirmar estos estudios utilizando un método diferente y sugieren que los datos sintéticos se pueden utilizar como una aproximación razonable para identificar tendencias en ensayos clínicos.

La aplicación de la tecnología de datos sintéticos a los ensayos clínicos supondría un importante avance científico y un ahorro de costes significativo al permitir generar pacientes sintéticos tanto para el brazo activo como para el brazo control de los ensayos. En el brazo activo permitiría generar modelos predictivos para identificar tendencias en fases tempranas de los ensayos que permitan a sus promotores tomar decisiones fundamentadas. Cuando las agencias reguladoras admitan los brazos de control sintéticos, esto permitirá reducir el número de pacientes control entre un 20-70% con el consiguiente ahorro de costos. Además, permitiría generar pacientes sintéticos fiables que complementan a los datos reales para un mejor aprendizaje de los algoritmos que generan gemelos digitales de enfermedades. Los gemelos digitales de enfermedades son modelos de simulación que permiten predecir como un paciente podría responder a un tratamiento y de esta forma seleccionar el mejor tratamiento personalizado para cada paciente. Finalmente permitiría generar pacientes sintéticos de enfermedades raras para mejorar la comprensión de estas patologías.

Conclusión

De los distintos algoritmos analizados, TVAE es el más adecuado y supone un excelente punto de partida para generar datos sintéticos a partir de los datos tabulares de ensayos clínicos. Los datos sintéticos permiten obtener conclusiones similares a las que ofrecen los datos reales en el análisis de los resultados de los ensayos clínicos.

Agradecimientos

Financiado por la Unión Europea – Next Generation EU. Inari agradece a la Comunidad de Madrid la concesión de ayudas del programa Investigo.

Bibliografía

1. Beaulieu-Jones, B.K., Wu, Z.S., Williams, C., Lee, R., Bhavnani, S.P., Byrd, J.B., and Greene, C.S. 2019. Privacy-preserving generative deep neural networks support clinical data sharing. Circ Cardiovasc Qual Outcomes, 12 (7), e005122. doi: 10.1161/CIRCOUTCOMES.118.005122.

2. Benaim, A.R., Almog, R., Gorelik, Y., Hochberg, I., Nassar, L., Mashiach, T., Khamaisi, M., Lurie, Y., Azzam, Z.S., Khoury, J., Kurnik, D., and Beyar, R. 2020. Analyzing medical research results based on synthetic data and their relation to real data results: systematic comparison from five observational studies. JMIR Med Inform, 8 (2), e16492. doi: 10.2196/16492.

3. Guillaudeux, M., Rousseau, O., Petot, J., Bennis, Z., Dein, C.A., Goronflot, T., Vince, N., Limou, S., Karakachoff, M., Wargny, M., and Gourraud, P.A. 2023. Patient-centric synthetic data generation, no reason to risk reidentification in biomedical data analysis. npj Digit Med, 6 (1), 37. doi: 10.1038/s41746-023-00771-5.

4. Azizi, Z., Zheng, C., Mosquera, L., Pilote, L., and Emam K.E. 2021. Can synthetic data be a proxy for real clinical trial data? A validation study. BMJ Open,11, e043497. doi:10.1136/ bmjopen-2020-043497.

5. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. 2014. Generative adversarial nets. Advances in Neural Information Processing Systems, 27, 2672–2680.

6. Zhavoronkov, A., Vanhaelen, Q., and Oprea, T.I. 2020. Will artificial intelligence for drug discovery impact clinical pharmacology?. Clin Phamacol Ther, 107 (4), 780-785.

7. Krenmayr, L., Frank, R., Drobig, C., Braungart, M., Seidel, J., Schaudt, D., Schwerin, R., and Stucke-Straub, K. 2022 GANerAid: Realistic synthetic patient data for clinical trials. Informatics in Medicine Unlocked, 35, 101118.

8. Alberts, A.R., Sinicrope, F.A., and Grothey, A. 2005. N0147: A randomized phase III trial of Oxaliplatin plus 5-Fluorouracil/Leucovorin with or without Cetuximab after curative resection of stage III colon cancer. Clinical Colorectal Cancer, 5 (3), 211-213.

9. Dahdaleh, F.S., Sherman, S.K., Poli, E.C., Vigneswaran, J., Polite, B.N., Sharma, M.R., Catenacci, D.V., Maron, S.B., and Turaga, K.K. 2018. Obstruction predicts worse long-term outcomes in stage III colon cancer: a secondary analysis of the N0147 trial. Surgery, 164, 1223–1229.

10. Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. 2019. Modeling Tabular data using Conditional GAN. Proc. of Advances in Neural Information Processing Systems, doi: 10.48550/arXiv.1907.00503.