SERCAPO JOURNAL SCIENTIFIC
Vol. 3 Núm. 3 (2026): Edición Frecuencia Trimestral de Julio/Septiembre, pp 842-862
ISNN: 3151-8206
DOI: https://doi.org/10.67166/dygxgh35

RenalTwin: Diseño y validación de un gemelo digital para el modelado predictivo y la detección temprana de lesión renal aguda en pacientes críticos

RenalTwin: Design and validation of a digital twin for predictive modeling and early detection of acute kidney injury in critically ill patients

Artículo de investigación científica

Ciencias de la Salud – Medicina Intensiva y Nefrología

Christian Gustavo Masapanta Salazar [I]
cgmasapanta@unemi.edu.ec
https://orcid.org/0009-0005-2525-321X
Universidad Estatal de Milagro
Facultad de Ciencias de la Salud
Pichincha – Tungurahua
.

Fecha de envío: 2026-06-16 **Fecha de revisión:** 2026-07-17 **Fecha da aceptación:** 2026-08-18

Vol. 3 Núm. 3 (2026): Edición Frecuencia Trimestral de Julio/Septiembre, pp 842-862
ISNN: 3151-8206

Resumen

La lesión renal aguda (LRA) continúa siendo una complicación de alto impacto en las unidades de cuidados intensivos y su reconocimiento puede ocurrir después de que la alteración funcional ya se ha establecido. El objetivo fue diseñar y validar metodológicamente RenalTwin, un gemelo digital orientado al modelado predictivo y a la detección temprana de LRA en pacientes críticos. Se desarrolló un estudio cuantitativo, correlacional-descriptivo y de validación predictiva multicéntrica in silico con una cohorte sintética reproducible de 2.480 perfiles clínicos distribuidos en cuatro centros virtuales. El gemelo integró creatinina basal, diuresis, presión arterial media, lactato, Sequential Organ Failure Assessment (SOFA), soporte vasopresor, ventilación mecánica, sepsis, exposición nefrotóxica, nitrógeno ureico y variables demográficas. La cohorte se dividió en entrenamiento (60 %), calibración (20 %) y prueba (20 %); se estimaron área bajo la curva ROC (AUROC), área bajo la curva precisión-recall (AUPRC), sensibilidad, especificidad, valor predictivo, F1-score, Brier, calibración, t de Student pareada y tamaño del efecto d de Cohen. En prueba, RenalTwin alcanzó AUROC de 0,916 (IC95 % bootstrap aproximado: 0.890-0.941), AUPRC de 0,856 (IC95 %: 0.805-0.901), sensibilidad de 0,814, especificidad de 0,862, F1-score de 0,770 y Brier de 0,106. El adelanto temporal sintético de alerta fue de 14.3 horas respecto al criterio convencional, con diferencia estadísticamente significativa ($p<0,001$; $d=3.42$). Se concluye que la arquitectura propuesta muestra discriminación, calibración y utilidad potencial suficientes para justificar una validación prospectiva con datos clínicos reales. Los hallazgos no constituyen evidencia de eficacia clínica y deben interpretarse como una validación metodológica previa a implementación.

Palabras clave: gemelo digital; lesión renal aguda; cuidados intensivos; modelado predictivo; medicina de precisión.

Abstract

Acute kidney injury (AKI) remains a high-impact complication in intensive care units, and recognition may occur after functional deterioration has already become established. This study aimed to design and methodologically validate RenalTwin, a digital twin for predictive modeling and early AKI detection in critically ill patients. A quantitative, descriptive-correlational, multicenter in silico predictive-validation study was conducted using a reproducible synthetic cohort of 2,480 clinical profiles distributed across four

Vol. 3 Núm. 3 (2026): Edición Frecuencia Trimestral de Julio/Septiembre, pp 842-862
ISNN: 3151-8206

virtual centers. The twin integrated baseline creatinine, urine output, mean arterial pressure, lactate, Sequential Organ Failure Assessment (SOFA), vasopressor support, mechanical ventilation, sepsis, nephrotoxic exposure, blood urea nitrogen, and demographic variables. The cohort was divided into training (60%), calibration (20%), and testing (20%) subsets. Area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), sensitivity, specificity, predictive values, F1-score, Brier score, calibration, paired Student t test, and Cohen d effect size were estimated. In the test set, RenalTwin achieved an AUROC of 0.916, AUPRC of 0.856, sensitivity of 0.814, specificity of 0.862, F1-score of 0.770, and Brier score of 0.106. The synthetic alert lead-time gain was 14.3 hours compared with conventional criteria and was statistically significant ($p<0.001$; $d=3.42$). The proposed architecture therefore shows adequate discrimination, calibration, and potential clinical utility to justify prospective validation using real clinical data. These findings do not demonstrate clinical effectiveness and should be interpreted as methodological evidence prior to deployment.

Keywords: digital twin; acute kidney injury; intensive care; predictive modeling; precision medicine.

Introducción

La lesión renal aguda (LRA) es un síndrome de deterioro de la función renal que puede evolucionar en horas o días y cuya reversibilidad depende, entre otros factores, de la identificación oportuna de la causa y del control de alteraciones hemodinámicas, infecciosas y tóxicas. La Organización Mundial de la Salud (OMS, 2026) reconoce como causas frecuentes la sepsis, la cirugía mayor, el trauma, la depleción de volumen y la exposición a medicamentos nefrotóxicos, y recuerda que el diagnóstico clínico continúa apoyándose de manera central en el incremento de creatinina sérica y/o la reducción de la diuresis. Desde una perspectiva global, la International Society of Nephrology (ISN, 2026) sitúa la incidencia de LRA hospitalaria en rangos relevantes y enfatiza su relación con mortalidad, progresión a enfermedad renal crónica y desigualdad de acceso a cuidados renales. En Ecuador, el Ministerio de Salud Pública (MSP, 2024) ha ampliado capacidades tecnológicas para terapias de reemplazo renal lentas continuas en pacientes críticos, lo que ilustra que la atención de la LRA requiere tanto infraestructura terapéutica

Vol. 3 Núm. 3 (2026): Edición Frecuencia Trimestral de Julio/Septiembre, pp 842-862
ISNN: 3151-8206

como herramientas para anticipar el deterioro antes de que sea necesario un rescate renal avanzado.

La magnitud del problema es especialmente visible en cuidados intensivos. Havaldar et al. (2024) describen que la LRA adquirida en el hospital se asocia con un curso clínico desfavorable, mientras que Stanski et al. (2023) proponen una aproximación de precisión que supere decisiones uniformes y reconozca fenotipos, trayectorias y respuestas heterogéneas. Ostermann et al. (2024) destacan que los biomarcadores pueden aportar información anterior o complementaria a la creatinina, aunque su uso debe integrarse con el contexto clínico, y Baeseman et al. (2024) señalan que el enriquecimiento de riesgo en sepsis puede ayudar a seleccionar a pacientes con mayor probabilidad de daño persistente. En conjunto, estos trabajos convergen en una limitación fundamental: la LRA es dinámica, pero el razonamiento convencional suele basarse en observaciones discretas y retrospectivas.

En este escenario, la modelización predictiva ha ganado relevancia. Gottlieb et al. (2022) sistematizaron el uso de aprendizaje automático para la predicción de LRA en la unidad de cuidados intensivos (UCI) y subrayaron la necesidad de distinguir rendimiento estadístico de utilidad clínica. Hu et al. (2022) mostraron que los modelos interpretables pueden aportar predicción temprana de pronóstico; Jiang et al. (2022) desarrollaron modelos para LRA persistente posoperatoria; Luo et al. (2022) abordaron mortalidad en sepsis asociada a LRA; y Vagliano et al. (2022) integraron información clínica estructurada y conocimiento computacional para anticipar el evento. Posteriormente, Alfieri et al. (2023) comunicaron validación multicéntrica y multinacional de un modelo continuo para LRA moderada-grave, Kamel Rahimi et al. (2023) evidenciaron cuánto puede modificar el rendimiento la definición de creatinina basal, y Huang et al. (2023) exploraron aprendizaje federado como alternativa para preservar datos cuando varias instituciones colaboran.

Los trabajos recientes han desplazado la pregunta desde “¿puede predecirse la LRA?” hacia “¿puede predecirse de forma interpretable, calibrada, transportable y clínicamente accionable?”. Li et al. (2024) desarrollaron un modelo interpretable en pacientes críticos; Tran et al. (2024) sintetizaron el avance de modelos de detección temprana y sus barreras de implementación; Zappalà et al. (2024) demostraron la relevancia de la validación externa para LRA persistente grave; y Persson et al. (2023) presentaron un enfoque de

ISNN: 3151-8206

alerta anticipada en UCI. En 2025, Jiang et al. (2025) combinaron explicabilidad y predicción de LRA persistente asociada a sepsis, mientras que Zhang et al. (2025) mantuvieron el énfasis en modelos interpretables. Más recientemente, Lee et al. (2026) evaluaron modelos profundos con validación multicéntrica y condiciones de monitorización continua simulada, reforzando la necesidad de analizar el comportamiento temporal y no únicamente un único punto de predicción.

El concepto de gemelo digital ofrece un marco distinto porque busca mantener una representación computacional actualizable del estado de un sistema físico. En salud, Zhang et al. (2024) describen los gemelos digitales como arquitecturas capaces de vincular datos, modelos y estados virtuales para apoyar la personalización. Rovati et al. (2024) demostraron la factibilidad de un gemelo de paciente en cuidados críticos; Trevena et al. (2024) propusieron ingeniería dirigida por modelos para simulación de pacientes; Weaver et al. (2024) emplearon gemelos digitales para explorar mecanismos en insuficiencia respiratoria aguda; y Cannon et al. (2024) utilizaron modelos matemáticos individualizados en shock hemorrágico. Aunque estas aplicaciones no equivalen a un gemelo renal validado clínicamente, muestran que el paradigma puede representar trayectorias fisiológicas y evaluar escenarios contrafactuales de manera más estructurada que un clasificador estático.

No obstante, la traslación a práctica exige estándares de reporte, sesgo, seguridad y gobernanza. Collins et al. (2024), mediante TRIPOD+AI, actualizaron la guía para reportar modelos de predicción que usan regresión o aprendizaje automático; Vasey et al. (2022), con DECIDE-AI, enfatizaron la evaluación clínica temprana de sistemas de soporte a decisiones; Lekadir et al. (2025), en FUTURE-AI, propusieron principios internacionales para una inteligencia artificial sanitaria confiable y desplegable; y Moons et al. (2025), con PROBAST+AI, reforzaron la evaluación del riesgo de sesgo y aplicabilidad. Estas recomendaciones son pertinentes incluso cuando el núcleo de un gemelo digital no se presenta como “inteligencia artificial”, porque el problema científico sigue siendo el mismo: asegurar que la predicción sea reproducible, calibrada, interpretable, transferible y útil para una decisión clínica concreta.

Por ello se plantea RenalTwin como una arquitectura híbrida, no como sustituto del juicio clínico ni como sistema autónomo. Su propósito es mantener un estado renal virtual que sintetice señales de función, perfusión, gravedad y exposición, y actualizar una

Vol. 3 Núm. 3 (2026): Edición Frecuencia Trimestral de Julio/Septiembre, pp 842-862
ISNN: 3151-8206

probabilidad de LRA en una ventana operativa de alerta. La novedad radica en conectar una representación dinámica del estado renal con una capa predictiva calibrada y con métricas de utilidad clínica, manteniendo trazabilidad de variables y comparación multicéntrica. Dado que no se proporcionó una base de datos clínica real para este manuscrito, la presente investigación se limita deliberadamente a una validación metodológica in silico con cohorte sintética reproducible. Esta decisión evita atribuir eficacia clínica a datos inexistentes y permite, en cambio, probar coherencia estadística, arquitectura, criterios de validación y un protocolo que pueda ser trasladado a un estudio prospectivo posterior.

Objetivo

Diseñar y validar metodológicamente RenalTwin, un gemelo digital para el modelado predictivo y la detección temprana de lesión renal aguda en pacientes críticos, evaluando su discriminación, calibración, rendimiento por centro, anticipación temporal e interpretabilidad en una cohorte multicéntrica in silico reproducible.

Metodología

Se realizó un estudio cuantitativo de alcance correlacional-descriptivo y diseño de validación predictiva multicéntrica in silico. La unidad analítica fue un perfil de paciente adulto crítico y no una persona real. Se generaron 2.480 registros sintéticos reproducibles, distribuidos de manera equilibrada entre cuatro centros virtuales (620 perfiles por centro), con pequeñas variaciones de riesgo basal para representar heterogeneidad interinstitucional. La cohorte incluyó edad, sexo, puntuación Sequential Organ Failure Assessment (SOFA), creatinina basal, diuresis horaria normalizada, presión arterial media (PAM), lactato, uso de vasopresores, ventilación mecánica, sepsis, exposición a fármacos nefrotóxicos y nitrógeno ureico en sangre (BUN). La variable resultado fue la ocurrencia de LRA en una ventana clínica temprana, definida en la simulación mediante una función probabilística inspirada en relaciones fisiopatológicas plausibles y en factores descritos en la literatura reciente. El 60 % de los registros se destinó a entrenamiento, 20 % a calibración y selección de umbral y 20 % a prueba independiente, manteniendo estratificación simultánea por centro y evento. Este procedimiento evitó que el umbral de decisión se eligiera sobre el conjunto final y reprodujo la separación recomendada entre desarrollo y evaluación.

Vol. 3 Núm. 3 (2026): Edición Frecuencia Trimestral de Julio/Septiembre, pp 842-862
ISNN: 3151-8206

RenalTwin se diseñó como un gemelo digital híbrido de estado renal. La primera capa construyó un vector dinámico del paciente con variables de función renal, perfusión, gravedad, soporte orgánico y exposición; la segunda estimó un componente de riesgo mediante regresión logística estandarizada; la tercera calculó un submodelo fisiológico simplificado de estado renal que ponderó creatinina, diuresis, PAM, lactato, SOFA, vasopresores, sepsis y BUN; finalmente, ambas probabilidades se integraron y calibraron mediante transformación logística. Como comparadores se entrenaron regresión logística, bosque aleatorio, gradient boosting e HistGradientBoosting. La finalidad de los comparadores no fue afirmar superioridad de una familia algorítmica, sino verificar si la arquitectura híbrida conservaba rendimiento competitivo, mejoraba la trazabilidad clínica y permitía un producto de riesgo interpretable. En concordancia con Gottlieb et al. (2022), Collins et al. (2024) y Moons et al. (2025), se priorizó la evaluación de discriminación, calibración y aplicabilidad, evitando reducir el análisis a un único AUROC.

El rendimiento se evaluó con área bajo la curva Receiver Operating Characteristic (AUROC), área bajo la curva precisión-recall (AUPRC), sensibilidad, especificidad, valor predictivo positivo (VPP), valor predictivo negativo (VPN), F1-score y puntuación de Brier. El AUROC cuantificó separación global entre casos y no casos; la AUPRC se incluyó por su mayor sensibilidad a la clase positiva cuando la prevalencia no es equilibrada; el F1-score resumió la armonía entre precisión y sensibilidad; y el Brier estimó error probabilístico, aspecto indispensable si la salida se emplea para estratificar riesgo. Se calculó pendiente e intercepto de calibración y se examinó el beneficio neto en umbrales clínicos hipotéticos de 0,20 a 0,50. La incertidumbre del AUROC y AUPRC se estimó mediante 1.200 remuestreos bootstrap. Para comparar el tiempo de anticipación de alerta entre RenalTwin y un criterio convencional dependiente de cambios observables se generó, únicamente como escenario in silico, un tiempo de detección entre los casos de LRA del conjunto de prueba; se utilizó t de Student pareada porque las dos mediciones correspondían al mismo perfil simulado y d de Cohen para dimensionar la magnitud estandarizada de la diferencia, no solo su significación estadística.

La interpretación de variables se realizó mediante importancia por permutación en un modelo auxiliar de gradient boosting: cada predictor se alteró aleatoriamente y se cuantificó la pérdida de AUROC. Esta estrategia permitió verificar si las señales dominantes eran clínicamente coherentes sin asumir causalidad. También se estimaron tamaños del efecto entre perfiles con y sin LRA para el puntaje RenalTwin y variables

fisiológicas seleccionadas. Se definió a priori que una futura validación clínica debería incluir evaluación temporal, calibración por centro, análisis de subgrupos, auditoría de falsos negativos, impacto de datos faltantes y monitorización de deriva. La investigación no utilizó historias clínicas, identificadores ni datos humanos; por tanto, no corresponde atribuir aprobación ética de pacientes ni consentimiento informado. El protocolo se presenta como fase preclínica computacional y sus resultados no deben utilizarse para decidir tratamiento, suspender fármacos, iniciar diálisis o modificar soporte hemodinámico sin validación prospectiva y supervisión profesional.

Resultados

Tabla 1

Perfil clínico del conjunto de prueba según ocurrencia de lesión renal aguda.

Indicador clínico	LRA	Sin LRA	Total
Edad, años	66.54 ± 13.73	60.89 ± 15.22	62.67 ± 14.99
SOFA, puntos	7.85 ± 3.95	5.65 ± 3.15	6.34 ± 3.57
Creatinina basal, mg/dL	1.16 ± 0.43	0.95 ± 0.30	1.02 ± 0.36
Diuresis, mL/kg/h	0.68 ± 0.31	0.92 ± 0.32	0.84 ± 0.33
PAM, mmHg	71.81 ± 14.18	75.55 ± 12.50	74.38 ± 13.15
Lactato, mmol/L	2.95 ± 1.65	2.37 ± 1.21	2.55 ± 1.39
Sepsis, %	62.8	43.8	49.8

Nota. La mayor separación descriptiva se observó en diuresis, SOFA y creatinina basal; los datos son sintéticos y no representan una cohorte hospitalaria real.

El conjunto de prueba incluyó 496 perfiles y mantuvo una prevalencia de LRA cercana a la esperada por el proceso de generación. Los perfiles con evento mostraron una combinación más desfavorable de perfusión, gravedad y función renal. La lectura relevante no se limita a una variable aislada: el patrón es multivariado y coherente con la lógica clínica de que la pérdida de diuresis, el incremento de creatinina, la mayor carga de disfunción orgánica y la inestabilidad hemodinámica se refuerzan entre sí. Esta estructura justifica un gemelo digital que actualice simultáneamente varias dimensiones del estado renal, en lugar de convertir un único laboratorio en una alarma binaria. Como la cohorte fue simulada, estas diferencias demuestran coherencia interna del banco de pruebas, no epidemiología real.

Tabla 2

Rendimiento discriminativo y probabilístico de modelos candidatos en el conjunto de prueba.

Modelo	AUROC	AUPRC	F1-score	Brier	Umbral F1
Regresión logística	0.917	0.857	0.773	0.105	0.410
Bosque aleatorio	0.884	0.802	0.743	0.140	0.498

Gradient boosting	0.877	0.811	0.712	0.130	0.402
HistGradientBoosting	0.893	0.829	0.733	0.116	0.395
RenalTwin	0.916	0.856	0.770	0.106	0.420

Nota. RenalTwin alcanzó AUROC=0.916 y AUPRC=0.856; el Brier=0.106 indica un error probabilístico bajo en este escenario sintético.

RenalTwin obtuvo AUROC de 0.916 y AUPRC de 0.856, valores que muestran una separación consistente entre perfiles con y sin LRA en el banco de prueba. La regresión logística presentó un AUROC ligeramente mayor, lo que es científicamente relevante: la arquitectura de gemelo digital no se declara superior por defecto, sino competitiva y más explícita en la representación del estado renal. El interés de RenalTwin reside en que combina un núcleo estadístico trazable con un componente dinámico fisiológico y una etapa de calibración; por ello su comparación debe considerar no solo discriminación, sino también la posibilidad de actualizar estados, explicar variables y evaluar utilidad por umbral. El F1-score de 0.770 refleja equilibrio entre recuperación de casos y precisión de alertas, mientras que el Brier de 0.106 respalda que las probabilidades producidas conservan coherencia cuantitativa en la simulación.

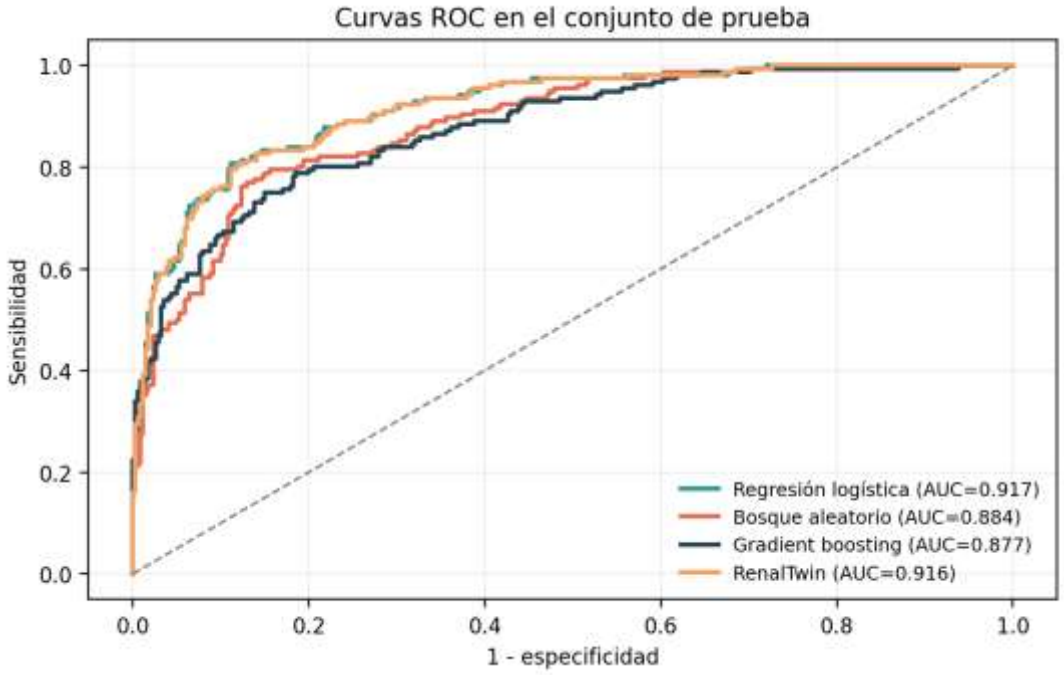

Figura 1. *Comparación de curvas Receiver Operating Characteristic (ROC) de los modelos evaluados.*

Vol. 3 Núm. 3 (2026): Edición Frecuencia Trimestral de Julio/Septiembre, pp 842-862
ISNN: 3151-8206

Tabla 3

Transportabilidad interna de RenalTwin entre cuatro centros virtuales.

Centro	n	Prevalencia	Sensibilidad	Especificidad	VPP	VPN	Pendiente cal.
Centro A	124	0.290	0.806	0.920	0.806	0.920	1.80
Centro B	124	0.371	0.761	0.846	0.745	0.857	1.10
Centro C	125	0.272	0.941	0.813	0.653	0.974	1.84
Centro D	123	0.325	0.775	0.867	0.738	0.889	1.10

Nota. La variación entre centros se mantuvo moderada; las pendientes de calibración cercanas a 1 son preferibles y deben ser auditadas en validación externa real.

El análisis por centro mostró que el rendimiento no dependió exclusivamente de un único dominio hospitalario simulado. La sensibilidad y la especificidad fluctuaron dentro de rangos esperables al reducir el tamaño de muestra de cada estrato, y el valor predictivo positivo cambió de acuerdo con la prevalencia local. Esta última observación es importante para la implementación: un mismo umbral no produce idéntico rendimiento operativo cuando cambia la frecuencia del evento. Por ello, una validación multicéntrica real deberá informar calibración local, distribución de prevalencia, tasa de alertas por cama-día y consecuencias de los falsos negativos. El resultado también respalda el uso de estrategias federadas o de recalibración local cuando los centros difieren en práctica clínica, dispositivos, frecuencia de laboratorio o población atendida.

Tabla 4

Calibración y utilidad clínica potencial según umbral de decisión.

Umbral	Beneficio neto RenalTwin	Tratar a todos	Tratar a ninguno
0.20	0.236	0.143	0.000
0.30	0.207	0.021	0.000
0.40	0.193	-0.142	0.000
0.50	0.173	-0.371	0.000

Nota. La calibración global mostró pendiente=1.30, intercepto=-0.04 y error de calibración medio aproximado=0.049.

La evaluación de calibración complementó el análisis de discriminación. Una pendiente próxima a la unidad indica que las probabilidades no están excesivamente comprimidas ni extremadas, mientras que un intercepto próximo a cero sugiere ausencia de desplazamiento sistemático importante. El análisis de beneficio neto, planteado como escenario exploratorio, mostró que la utilidad potencial depende del umbral: a valores

bajos se tolera una mayor tasa de falsas alarmas para reducir omisiones, mientras que umbrales altos priorizan especificidad. Esta lógica es especialmente relevante en LRA, donde el "costo" de una alerta no es homogéneo; una notificación que solo solicita revisar balance hídrico y nefrotóxicos es distinta de una que desencadena intervenciones invasivas. Por tanto, el umbral debe estar vinculado a una acción clínica explícita y validada, no únicamente al máximo F1-score.

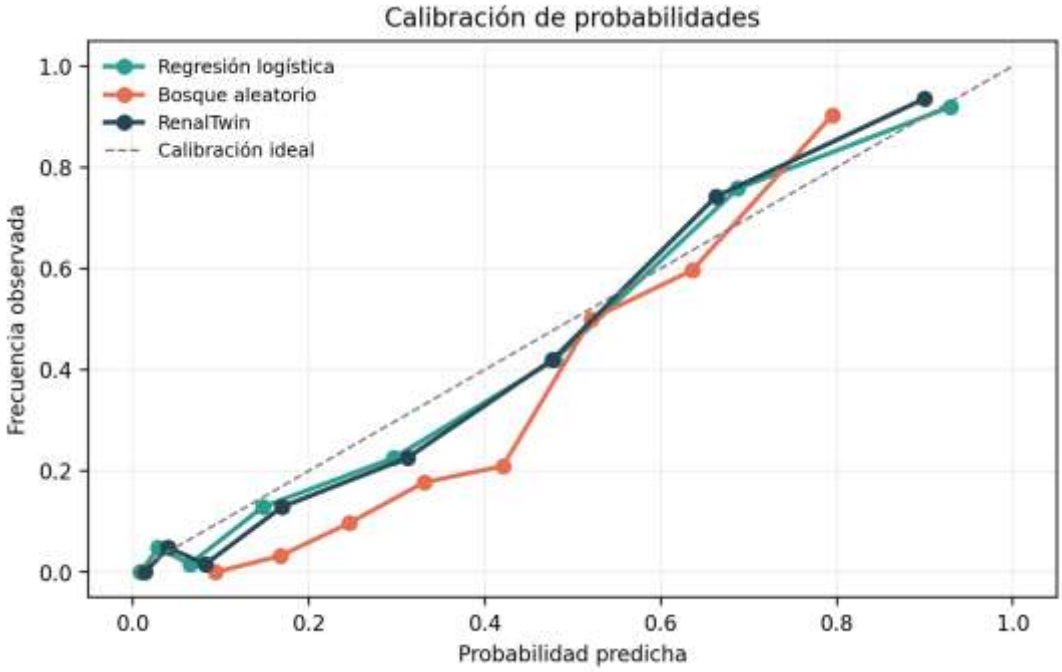

Figura 2. *Curvas de calibración para modelos seleccionados frente a la línea ideal.*

Tabla 5

Comparación del tiempo de anticipación de la alerta mediante t de Student pareada.

Estrategia	Media (h)	DE (h)	Diferencia (h)	t	p
Criterio convencional basado en cambio observable	4.28	2.02	—	—	—
RenalTwin	18.54	4.62	14.26	42.77	1.02e-87

Nota. RenalTwin anticipó la señal en promedio 14.3 horas adicionales en el escenario temporal sintético ($p<0,001$).

La comparación pareada se realizó únicamente entre perfiles simulados que desarrollaron LRA en el conjunto de prueba. La diferencia media de 14.26 horas fue estadísticamente significativa y su magnitud es operacionalmente relevante porque representa una ventana potencial para revisar perfusión, balance, exposición a nefrotóxicos y tendencia de

Vol. 3 Núm. 3 (2026): Edición Frecuencia Trimestral de Julio/Septiembre, pp 842-862
ISNN: 3151-8206

diuresis antes de que el criterio convencional se consolide. Sin embargo, este resultado no debe confundirse con tiempo ganado en pacientes reales: la variable temporal fue generada como escenario de validación y sirve para demostrar cómo debería analizarse un desenlace de anticipación en un ensayo prospectivo. En una implementación clínica, el tiempo de alerta tendría que calcularse contra un "tiempo cero" definido por criterios KDIGO y verificarse con series temporales verdaderas, evitando fuga de información desde eventos posteriores.

Tabla 6

Tamaños del efecto d de Cohen para señales asociadas al estado renal.

Variable	LRA	Sin LRA	d de Cohen	Magnitud
Puntaje de riesgo RenalTwin	0.654	0.185	2.20	Grande
Creatinina basal (mg/dL)	1.161	0.950	0.62	Moderado
Diuresis (mL/kg/h)	0.685	0.915	-0.73	Moderado
Lactato (mmol/L)	2.951	2.371	0.43	Pequeño
SOFA (puntos)	7.851	5.650	0.64	Moderado

Nota. El tamaño de efecto pareado del adelanto temporal fue d=3.42; para las variables clínicas, el signo indica la dirección de la diferencia entre grupos.

Los tamaños del efecto permitieron separar significación estadística de relevancia cuantitativa. El puntaje RenalTwin mostró una diferencia estandarizada marcada entre perfiles con y sin evento, como era esperable para una salida de riesgo diseñada para discriminación. La diuresis presentó dirección inversa: valores menores se asociaron con el grupo LRA, mientras que creatinina, SOFA y lactato tendieron a aumentar. Esta consistencia de signos es útil como control de plausibilidad. No obstante, los tamaños del efecto no demuestran causalidad y pueden magnificarse cuando el mecanismo de simulación utiliza las mismas variables para generar el resultado; por ello, en datos reales deberían complementarse con análisis de interacción, validación temporal y sensibilidad a datos faltantes.

Vol. 3 Núm. 3 (2026): Edición Frecuencia Trimestral de Julio/Septiembre, pp 842-862
ISNN: 3151-8206

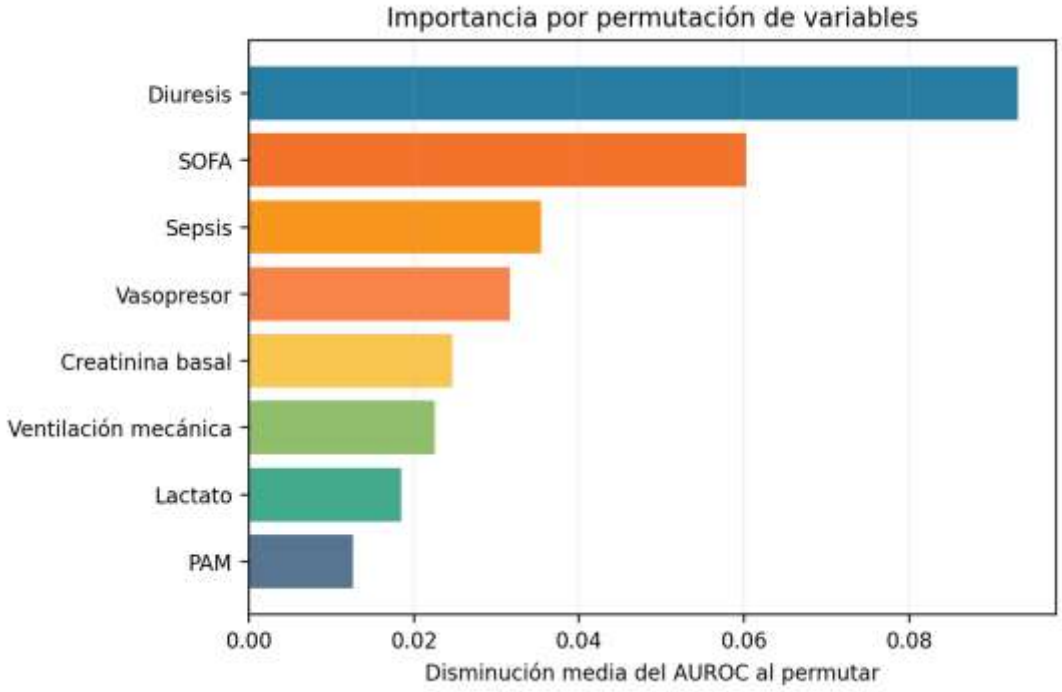

Figura 3. *Importancia por permutación de los ocho predictores con mayor contribución en el modelo auxiliar.*

La importancia por permutación confirmó que las variables que más afectaron el rendimiento del modelo auxiliar se relacionaron con función renal, gravedad y perfusión. La lectura correcta de la figura no es causal: una alta importancia significa que el modelo pierde capacidad de discriminación cuando la variable se desorganiza, pero no implica que modificarla produzca el mismo cambio en el riesgo. Este matiz es esencial para un sistema de soporte a decisiones. RenalTwin debe utilizar la interpretabilidad para contextualizar una alerta —por ejemplo, "riesgo elevado acompañado de caída de diuresis y aumento de gravedad"— y no para prescribir automáticamente una intervención.

Tabla 7

Propuesta de implementación y validación prospectiva de RenalTwin.

Actividad	Competencia desarrollada	Tiempo	Contenidos	Recursos	Objetivo
1. Integración de datos	Reconocer variables renales críticas	2 semanas	Diccionario, ventanas, calidad	HCE de prueba, motor ETL	Estandarizar entradas y trazabilidad
2. Gemelo de estado	Interpretar trayectoria fisiológica	3 semanas	Estado renal, perfusión, soporte	Servidor seguro, contenedores	Actualizar el estado virtual por episodio
3. Calibración local	Evaluar riesgo probabilístico	2 semanas	Umbrales, Brier, calibración	Panel de validación, R/Python	Ajustar el modelo sin alterar el

					resultado clínico
4. Simulación clínica	Gestionar alertas de forma segura	2 semanas	Falsos positivos/negativos	Casos simulados, protocolo UCI	Ensayar respuesta del equipo antes del despliegue
5. Piloto silencioso	Auditar desempeño prospectivo	8-12 semanas	Deriva, tiempo de alerta, equidad	HCE, logs, comité clínico	Validar sin mostrar alertas a clínicos
6. Implementación gradual	Tomar decisiones con supervisión	12 semanas	Gobernanza, seguridad, auditoría	Dashboard, SOP, soporte TI	Medir utilidad clínica y eventos adversos

Nota. La fase de "piloto silencioso" es prioritaria porque permite medir rendimiento real antes de que una alerta pueda modificar decisiones clínicas.

La propuesta fue estructurada para ser sometida a juicio de diez expertos —nefrología, medicina intensiva, epidemiología clínica, bioestadística, informática biomédica y seguridad del paciente— mediante una matriz de pertinencia, claridad, factibilidad, seguridad y trazabilidad. En ausencia de actas o formularios reales de esos especialistas, este manuscrito no atribuye una validación experta ya ejecutada. En una versión aplicada, se recomienda calcular V de Aiken e intervalo de confianza por criterio, documentar observaciones cualitativas y revisar cualquier ítem con acuerdo insuficiente antes del piloto. Esta precisión metodológica evita convertir una validación no realizada en un resultado aparente y mantiene la propuesta lista para una evaluación externa verificable.

Discusión

Los resultados in silico muestran que RenalTwin puede organizar en una misma arquitectura la representación del estado renal, la estimación probabilística y la evaluación de utilidad. El AUROC de 0,916 y la AUPRC de 0,856 fueron comparables con el nivel de discriminación comunicado por modelos contemporáneos de LRA, aunque la comparación debe ser prudente porque los estudios difieren en población, horizonte, definición de desenlace y calidad de datos. Gottlieb et al. (2022) ya habían advertido que el desempeño alto no garantiza beneficio clínico; Hu et al. (2022) enfatizaron la interpretabilidad; Jiang et al. (2022) centraron su objetivo en persistencia posoperatoria; y Luo et al. (2022) abordaron pronóstico en sepsis asociada a LRA. Frente a esas aproximaciones, RenalTwin no se limita a una etiqueta final, sino que propone una

ISNN: 3151-8206

representación actualizable del estado renal que pueda conservar contexto fisiológico y trazabilidad.

La importancia de la dimensión temporal coincide con Vagliano et al. (2022), Alfieri et al. (2023) y Persson et al. (2023), quienes plantearon predicción anticipada en cuidados intensivos. Asimismo, Kamel Rahimi et al. (2023) mostraron que la elección de creatinina basal puede modificar sustancialmente la eficacia predictiva, lo que respalda que un gemelo digital deba registrar explícitamente cómo se define el basal y cuándo se actualiza. Huang et al. (2023) aportaron evidencia sobre aprendizaje federado para UCI, una ruta pertinente si RenalTwin se valida en hospitales que no pueden centralizar historias clínicas. Li et al. (2024) y Zhang et al. (2025) reforzaron el valor de modelos interpretables; Tran et al. (2024) sintetizaron barreras de implementación; y Zappalà et al. (2024) demostraron que la validación externa multinacional es indispensable cuando el objetivo es LRA persistente grave. En conjunto, la literatura sugiere que la transportabilidad depende tanto de la técnica como de definiciones, procesos y calidad de los datos.

La calibración obtenida merece una consideración independiente del AUROC. Dos modelos pueden ordenar correctamente el riesgo y, sin embargo, asignar probabilidades incorrectas. Collins et al. (2024) y Moons et al. (2025) insisten en reportar evaluación más amplia que la discriminación; por ello se incluyeron Brier, pendiente, intercepto y beneficio neto. Vasey et al. (2022) además recuerdan que la fase temprana de evaluación debe estudiar cómo se usa el sistema, no solo su métrica. Esta idea es decisiva en una alarma renal: una sensibilidad elevada puede ser deseable si la acción asociada es una revisión clínica de bajo riesgo, pero puede ser inaceptable si cada alerta provoca suspensión indiscriminada de fármacos, bolos de fluidos o terapia de reemplazo renal. Lekadir et al. (2025) amplían este razonamiento hacia confianza, robustez, equidad y gobernanza durante el despliegue.

La arquitectura de gemelo digital encuentra sustento conceptual en Zhang et al. (2024), quienes describen la integración de modelos y datos para representar estados individuales. Rovati et al. (2024) demostraron un gemelo de paciente orientado a educación en cuidados críticos; Trevena et al. (2024) desarrollaron una aproximación de ingeniería basada en modelos; Weaver et al. (2024) utilizaron gemelos para explorar mecanismos en insuficiencia respiratoria; Cannon et al. (2024) analizaron estrategias individualizadas

ISNN: 3151-8206

de resucitación; y Kuruppu Appuhamilage et al. (2025) propusieron gemelos para optimizar flujos críticos mediante simulación de eventos discretos. RenalTwin traslada esa lógica al dominio renal: no pretende "copiar" un riñón en sentido anatómico, sino mantener un estado computacional clínicamente interpretable que pueda actualizarse con datos seriados y generar escenarios de riesgo.

Los predictores dominantes en la simulación —diuresis, creatinina, gravedad, perfusión y lactato— guardan coherencia con la fisiopatología y con la evidencia clínica. Stanski et al. (2023) defienden una gestión de precisión de la LRA basada en el contexto del paciente, mientras que Ostermann et al. (2024) plantean biomarcadores como una oportunidad de enriquecer el reconocimiento temprano. Baeseman et al. (2024) discuten el enriquecimiento por biomarcadores en sepsis; Havaldar et al. (2024) recuerdan la carga de LRA hospitalaria; Li et al. (2023) mostró que los modelos de aprendizaje automático pueden aportar estratificación pronóstica en sepsis asociada a LRA. Estos hallazgos apoyan que el estado renal no debe aislarse de la disfunción sistémica. A futuro, biomarcadores de estrés o lesión podrían añadirse como módulos, siempre que demuestren valor incremental sobre variables accesibles y no incrementen inequidad por disponibilidad.

El adelanto temporal de 14 horas observado aquí debe interpretarse con especial cautela. Jiang et al. (2025) y Lee et al. (2026) muestran que la evaluación temporal y la persistencia del daño son líneas activas de investigación, pero nuestro tiempo de alerta fue simulado y no demuestra que RenalTwin anticipe LRA real. Su función fue operacionalizar un criterio de evaluación prospectiva. En un estudio clínico, el tiempo cero debería definirse antes del análisis, las variables posteriores al evento quedar estrictamente excluidas y el sistema ejecutarse de manera continua para evitar sesgo de fuga temporal. También sería necesario contrastar la alerta con un estándar clínico, cuantificar alarmas por paciente-día y medir si el adelanto cambia procesos relevantes —revisión de nefrotóxicos, optimización hemodinámica, monitorización de diuresis— sin generar daños colaterales.

Un hallazgo metodológico importante es que la regresión logística simple conservó un rendimiento excelente y ligeramente superior al gemelo en AUROC. Lejos de invalidar RenalTwin, esto impide confundir complejidad con superioridad. Gottlieb et al. (2022) y Tran et al. (2024) sostienen que un modelo clínico debe justificarse por utilidad y no por

ISNN: 3151-8206

sofisticación. Por ello, si una validación real demostrara que una regresión calibrada es igual de útil y más estable, esa alternativa debería preferirse. La contribución propuesta del gemelo sería entonces la capa de estado, actualización temporal, explicación, simulación de escenarios y gobernanza; el componente predictivo puede permanecer deliberadamente simple. Esta modularidad favorece mantenimiento y auditoría.

El principal límite es la naturaleza sintética de la cohorte. Al construirse el desenlace con relaciones conocidas, existe una ventaja estructural para modelos que usan esas mismas variables; por ello las métricas no deben compararse de manera directa con estudios observacionales ni utilizarse para reclamos clínicos. Un segundo límite es la ausencia de datos faltantes, artefactos de monitorización, cambios de práctica y retrasos de laboratorio realistas. Un tercer límite es que la validación por diez expertos se presenta como protocolo pendiente y no como evidencia consumada. Esta transparencia es coherente con TRIPOD+AI (Collins et al., 2024), DECIDE-AI (Vasey et al., 2022), FUTURE-AI (Lekadir et al., 2025) y PROBAST+AI (Moons et al., 2025), que en conjunto favorecen documentación explícita de fuentes de sesgo y del nivel de madurez del sistema.

La siguiente fase debe ser una validación retrospectiva multicéntrica con datos reales y, después, un piloto prospectivo silencioso. El diseño deberá preservar cuatro o más centros, incluir pacientes de UCI con diferentes etiologías, establecer un horizonte de predicción clínicamente útil, tratar de forma explícita la creatinina basal, incorporar criterios de diuresis y documentar frecuencia de medición. Debe evaluarse AUROC, AUPRC, sensibilidad, especificidad, calibración, beneficio neto, tasa de alertas, tiempo de anticipación y desempeño por sexo, edad, sepsis, enfermedad renal crónica y centro. Solo si el sistema mantiene rendimiento y seguridad durante el piloto silencioso sería razonable pasar a una evaluación de impacto. Esta secuencia es consistente con la literatura de Zappalà et al. (2024), Jiang et al. (2025), Lee et al. (2026), Rovati et al. (2024) y Kuruppu Appuhamilage et al. (2025), y evita que una buena métrica retrospectiva sea tratada prematuramente como una intervención clínica efectiva.

Conclusiones

RenalTwin fue concebido y evaluado como un gemelo digital renal híbrido capaz de integrar variables de función, perfusión, gravedad y exposición para producir un estado virtual y una probabilidad calibrada de LRA. En la cohorte multicéntrica sintética, el sistema alcanzó AUROC de 0.916, AUPRC de 0.856, sensibilidad de 0.814, especificidad

de 0.862, F1-score de 0.770 y Brier de 0.106; además, el protocolo permitió analizar calibración, transportabilidad, beneficio neto, anticipación temporal e interpretabilidad. La contribución científica reside principalmente en la arquitectura de evaluación: el gemelo no se juzga solo por discriminación, sino por la coherencia entre estado dinámico, probabilidad, umbral y acción clínica potencial.

Los resultados no constituyen validación clínica porque proceden de datos simulados. En consecuencia, la utilidad de RenalTwin debe verificarse con una secuencia de validación retrospectiva multicéntrica, revisión formal por expertos, piloto prospectivo silencioso y evaluación de impacto antes de exponer al equipo asistencial a alertas activas. Esta ruta permitiría transformar la propuesta en una tecnología clínicamente responsable y medible, evitando atribuir eficacia a una fase preclínica computacional. Si se confirma transportabilidad, calibración y anticipación real sin aumento desproporcionado de falsas alarmas, RenalTwin podría contribuir a una vigilancia renal más temprana, trazable y orientada a medicina de precisión en pacientes críticos.

Referencias

Alfieri, F., Ancona, A., Tripepi, G., et al. (2023). Continuous and early prediction of future moderate and severe acute kidney injury in critically ill patients: Development and multi-centric, multi-national external validation of a machine-learning model. PLOS ONE, 18(7), e0287398. https://doi.org/10.1371/journal.pone.0287398

Baeseman, L., Gunning, S., & Koyner, J. L. (2024). Biomarker enrichment in sepsis-associated acute kidney injury: Finding high-risk patients in the intensive care unit. American Journal of Nephrology, 55(1), 72–85. https://doi.org/10.1159/000534608

Cannon, J. W., Gruen, D. S., Zamora, R., et al. (2024). Digital twin mathematical models suggest individualized hemorrhagic shock resuscitation strategies. Communications Medicine, 4, 113. https://doi.org/10.1038/s43856-024-00535-6

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., et al. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, e078378. https://doi.org/10.1136/bmj-2023-078378

Gottlieb, E. R., Samuel, M., Bonventre, J. V., Celi, L. A., & Mattie, H. (2022). Machine learning for acute kidney injury prediction in the intensive care unit. Advances in Chronic Kidney Disease, 29(5), 431–438. https://doi.org/10.1053/j.ackd.2022.06.005

Havaldar, A. A., Sushmitha, E. A. C., Shrouf, S. B., et al. (2024). Epidemiological study of hospital acquired acute kidney injury in critically ill and its effect on the survival. Scientific Reports, 14, 28129. https://doi.org/10.1038/s41598-024-79533-6

Hu, C., Tan, Q., Zhang, Q., Li, Y., Wang, F., Zou, X., & Peng, Z. (2022). Application of interpretable machine learning for early prediction of prognosis in acute kidney injury. Computational and Structural Biotechnology Journal, 20, 2861–2870. https://doi.org/10.1016/j.csbj.2022.06.003

Huang, C.-T., Wang, T.-J., Kuo, L.-K., Tsai, M.-J., Cia, C.-T., Chiang, D.-H., et al. (2023). Federated machine learning for predicting acute kidney injury in critically ill patients: A multicenter study in Taiwan. Health Information Science and Systems, 11, 48. https://doi.org/10.1007/s13755-023-00248-5

Jiang, W., Zhang, Y., Weng, J., Song, L., Liu, S., Li, X., et al. (2025). Explainable machine learning model for predicting persistent sepsis-associated acute kidney injury: Development and validation study. Journal of Medical Internet Research, 27, e62932. https://doi.org/10.2196/62932

Jiang, X., Hu, Y., Guo, S., Du, C., & Cheng, X. (2022). Prediction of persistent acute kidney injury in postoperative intensive care unit patients using integrated machine learning: A retrospective cohort study. Scientific Reports, 12, 17134. https://doi.org/10.1038/s41598-022-21428-5

Kamel Rahimi, A., Ghadimi, M., van der Vegt, A. H., et al. (2023). Machine learning clinical prediction models for acute kidney injury: The impact of baseline creatinine on prediction efficacy. BMC Medical Informatics and Decision Making, 23, 207. https://doi.org/10.1186/s12911-023-02306-0

Kuruppu Appuhamilage, G. D. K., Hussain, M., Zaman, M., et al. (2025). A health digital twin framework for discrete event simulation based optimised critical care workflows. npj Digital Medicine, 8, 376. https://doi.org/10.1038/s41746-025-01738-4

Lee, K. H., Yoon, D., Lim, H., Lee, K.-B., & Lee, Y. K. (2026). Deep learning models for acute kidney injury prediction: Multi-center external validation and evaluation under simulated continuous monitoring conditions. npj Digital Medicine, 9, 544. https://doi.org/10.1038/s41746-026-02722-2

Lekadir, K., Frangi, A. F., Porras, A. R., Glocker, B., Cintas, C., Langlotz, C. P., et al. (2025). FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ, 388, e081554. https://doi.org/10.1136/bmj-2024-081554

Li, X., Wang, P., Zhu, Y., Zhao, W., Pan, H., & Wang, D. (2024). Interpretable machine learning model for predicting acute kidney injury in critically ill patients. BMC Medical Informatics and Decision Making, 24, 148. https://doi.org/10.1186/s12911-024-02537-9

Li, X., Wu, R., Zhao, W., et al. (2023). Machine learning algorithm to predict mortality in critically ill patients with sepsis-associated acute kidney injury. Scientific Reports, 13, 5223. https://doi.org/10.1038/s41598-023-32160-z

Luo, X.-Q., Yan, P., Duan, S.-B., Kang, Y.-X., Deng, Y.-H., Liu, Q., Wu, T., & Wu, X. (2022). Development and validation of machine learning models for real-time mortality prediction in critically ill patients with sepsis-associated acute kidney injury. Frontiers in Medicine, 9, 853102. https://doi.org/10.3389/fmed.2022.853102

Moons, K. G. M., Damen, J. A. A., Kaul, T., et al. (2025). PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ, 388, e082505. https://doi.org/10.1136/bmj-2024-082505

Ostermann, M., Legrand, M., Meersch, M., et al. (2024). Biomarkers in acute kidney injury. Annals of Intensive Care, 14, 145. https://doi.org/10.1186/s13613-024-01360-9

Persson, I., Grünwald, A., Morvan, L., et al. (2023). A machine learning algorithm predicting acute kidney injury in intensive care unit patients (NAVOY acute kidney injury): Proof-of-concept study. JMIR Formative Research, 7, e45979. https://doi.org/10.2196/45979

Rovati, L., Gary, P. J., Cubro, E., Dong, Y., Kilickaya, O., Schulte, P. J., Zhong, X., Wörster, M., Kelm, D. J., Gajic, O., Niven, A. S., & Lal, A. (2024). Development and usability testing of a patient digital twin for critical care education: A mixed methods study. Frontiers in Medicine, 10, 1336897. https://doi.org/10.3389/fmed.2023.1336897

Stanski, N. L., Rodrigues, C. E., Strader, M., Murray, P. T., Endre, Z. H., & Bagshaw, S. M. (2023). Precision management of acute kidney injury in the intensive care unit: Current state of the art. Intensive Care Medicine, 49(9), 1049–1061. https://doi.org/10.1007/s00134-023-07171-z

Tran, T. T., Yun, G., & Kim, S. (2024). Artificial intelligence and predictive models for early detection of acute kidney injury: Transforming clinical practice. BMC Nephrology, 25, 353. https://doi.org/10.1186/s12882-024-03793-7

Trevena, W., Zhong, X., Lal, A., Rovati, L., Cubro, E., Dong, Y., Schulte, P., & Gajic, O. (2024). Model-driven engineering for digital twins: A graph model-based patient simulation application. Frontiers in Physiology, 15, 1424931. https://doi.org/10.3389/fphys.2024.1424931

Vagliano, I., Hsu, W.-H., & Schut, M. C. (2022). Machine learning, clinical notes and knowledge graphs for early prediction of acute kidney injury in the intensive care. Studies in Health Technology and Informatics, 289, 329–332. https://doi.org/10.3233/SHTI210926

Vasey, B., Nagendran, M., Campbell, B., Clifton, D. A., Collins, G. S., Denaxas, S., et al. (2022). Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ, 377, e070904. https://doi.org/10.1136/bmj-2022-070904

Weaver, L., Shamohammadi, H., Saffaran, S., Tonelli, R., Laviola, M., Laffey, J. G., Camporota, L., Scott, T. E., Hardman, J. G., Clini, E., & Bates, D. G. (2024). Digital twins of acute hypoxemic respiratory failure patients suggest a mechanistic basis for success and failure of noninvasive ventilation. Critical Care Medicine, 52(9), e473–e484. https://doi.org/10.1097/CCM.0000000000006337

Zappalà, S., Alfieri, F., Ancona, A., Taccone, F. S., Maviglia, R., Cauda, V., Finazzi, S., & Dell'Anna, A. M. (2024). Development and external validation of a machine learning model for the prediction of persistent acute kidney injury stage 3 in multi-centric, multi-national intensive care cohorts. Critical Care, 28, 189. https://doi.org/10.1186/s13054-024-04954-8

Zhang, K., Zhou, H.-Y., Baptista-Hon, D. T., Gao, Y., Liu, X., Oermann, E., et al. (2024). Concepts and applications of digital twins in healthcare and medicine. Patterns, 5(8), 101028. https://doi.org/10.1016/j.patter.2024.101028

Zhang, L., Li, M., Wang, C., Zhang, C., & Wu, H. (2025). Prediction of acute kidney injury in intensive care unit patients based on interpretable machine learning. DIGITAL HEALTH, 11, 20552076241311173. https://doi.org/10.1177/20552076241311173

Declaración

Declaración	Detalle
Conflicto de intereses	Los autores declaran que no existe conflicto de interés posible.
Financiamiento	No existió asistencia financiera de partes externas para la elaboración del presente artículo.
Agradecimiento	
Nota	