Validación externa de modelos predictivos para progresión en la enfermedad renal crónica en población colombiana

Luis H Rojas, MSc; Santiago Gonzalez, BSc; Walberto Buelvas, MD; Angela Pereira-Morales, PhD

Instituciones participantes: Medisinú, S4L SAS

Fecha: 12-02- 2023

DESCRIPCIÓN DE LOS MODELOS DESARROLLADOS Y VALIDADOS INTERNAMENTE

eTFG

Desarrollado como parte de un enfoque de aprendizaje supervisado, este modelo se ha entrenado utilizando datos longitudinales de seguimiento de aproximadamente 50,000 pacientes atendidos entre los años 2017 y 2023 en un prestador especializado en el manejo de enfermedades crónicas.

El propósito principal de este modelo es prever el desenlace de la eTFG, un indicador crítico de la función renal. Entre los predictores utilizados se encuentran variables como sexo, edad, peso, creatinina sérica, niveles de lipoproteínas, hemoglobina glicosilada, diagnósticos médicos y uso de medicamentos específicos.

Para garantizar su robustez y generalización, el modelo ha sido validado internamente utilizando métodos de validación cruzada y Bootstrap, y se ha evaluado su desempeño mediante métricas clave como el Error Absoluto Medio (MAE), el Error Cuadrático Medio (MSE) y el Coeficiente de Determinación (R^2). Con un MAE de 1.6205, un MSE de 6.8027 y un R^2 de 0.9510, el modelo muestra una capacidad excelente para predecir la eTFG en pacientes con enfermedades crónicas.

Respecto al peso relativo de cada característica en el modelo, destacan la creatinina y la eTFG en línea de base y el peso como predictores más relevantes.

Progresión rápida

El modelo predictivo de progresión rápida ha sido diseñado con el propósito de identificar tempranamente a pacientes con un mayor riesgo de experimentar una aceleración en la progresión de la enfermedad renal en un lapso de 12 meses. Utilizando un enfoque de clasificación binaria, el modelo asigna a cada paciente a una de dos categorías: alto riesgo o bajo riesgo de progresión renal acelerada. Esta clasificación se realiza considerando un umbral de progresión acelerada establecido en 5 ml por año, y su interpretación se proyecta dentro de un horizonte temporal de un año desde la fecha de predicción.

Este modelo se ha desarrollado mediante el aprendizaje supervisado y ha sido entrenado utilizando datos longitudinales de seguimiento de pacientes atendidos entre 2017 y 2023 en un centro especializado en

enfermedades crónicas. El conjunto de datos de entrenamiento consta de 50,000 pacientes, con una estructura de 23 características por paciente.

Entre los predictores utilizados se encuentran variables como sexo, edad, peso basal y actual, creatinina sérica, perfiles lipídicos, diagnósticos médicos y el uso de medicamentos específicos relacionados con las enfermedades precursoras.

En cuanto al desempeño del modelo, se han obtenido métricas sólidas, incluyendo una precisión del 96%, una sensibilidad del 92%, una especificidad del 98% y un puntaje F1 del 94%. Estas métricas demuestran la capacidad del modelo para identificar de manera precisa a los pacientes en riesgo de progresión renal acelerada.

La población objetivo para este modelo incluye a pacientes en seguimiento por enfermedades precursoras como diabetes mellitus tipo 2, hipertensión arterial y enfermedad renal crónica. Se deben tener en cuenta ciertas limitaciones y consideraciones al interpretar los resultados, como la exclusión de pacientes en terapia de reemplazo renal y aquellos con enfermedades autoinmunes.

Cambio de estadio renal

El modelo predictivo tiene el propósito de identificar anticipadamente a pacientes con mayor riesgo de cambiar de estadio renal en un periodo de 12 meses. Utilizando clasificación multiclase, el modelo predice cinco categorías: ERC1, ERC2, ERC3a, ERC3b, y ERC4. Este modelo se basa en datos de pacientes atendidos entre 2017 y 2023 en una institución especializada en enfermedades crónicas.

La población objetivo para este modelo son pacientes con enfermedades precursoras como diabetes tipo 2, hipertensión arterial y enfermedad renal crónica, con edades comprendidas entre 18 y 80 años, excluyendo aquellos en terapia de reemplazo renal y con enfermedades autoinmunes.

Se han utilizado diversos predictores como sexo, edad, IMC, niveles de creatinina, lípidos séricos, medicación, y diagnósticos de comorbilidades. El modelo ha alcanzado una alta puntuación de 0.978 en el área bajo la curva ROC (AUC), lo que indica una excelente capacidad discriminativa. Además, para la clase 0, el modelo alcanza una precisión del 91%, con un recall del 59% y un F1-score de 0.71. Para la clase 1, los valores son aún más altos, con una precisión del 93%, un recall del 96% y un F1-score de 0.94. Similarmente, se observan buenos resultados para las clases restantes, con precisión, recall y F1-score que oscilan entre el 85% y el 89%. La precisión global del modelo, medida por el accuracy, es del 91%, lo que indica una capacidad general satisfactoria para clasificar correctamente las muestras. En promedio, las métricas macro y weighted avg muestran una precisión del 74% y 90%, respectivamente, lo que subraya la capacidad del modelo para generalizar y mantener un buen rendimiento en todas las clases de riesgo renal.

Se han empleado métodos de validación como la validación cruzada (cv=5) y la división de datos en conjuntos de entrenamiento y prueba (80% - 20%). Además, se realizó una optimización de hiperparámetros para mejorar el rendimiento del modelo.

Finalmente, la currva ROC multiclase, que varía desde 0.97 hasta 0.99, confirma la robustez del modelo en la clasificación de riesgo renal.

Relación albuminuria creatinuria (RAC)

El modelo predictivo RAC, diseñado para anticipar el riesgo de daño renal en pacientes en un lapso de 12 meses, se fundamenta en un enfoque de aprendizaje supervisado. Entrenado con datos de seguimiento de aproximadamente 50,000 pacientes entre 2017 y 2023, atendidos en un centro especializado en enfermedades crónicas, este modelo utiliza la relación albuminuria creatinuria (RAC) como variable de predicción. Su objetivo principal es prever el rango de RAC en el que se ubicará cada paciente al cabo de un año, facilitando así la identificación temprana de individuos con mayor riesgo renal.

Para lograr estas predicciones, el modelo emplea diversos predictores, entre ellos sexo, edad, peso, creatinina sérica, perfil lipídico, hemoglobina glicosilada, diagnósticos médicos y el uso de medicamentos específicos. Además, integra mediciones de tasa de filtración glomerular estimada (eTFG) tanto en la línea de base como en el seguimiento actual del paciente.

En cuanto a su desempeño, el modelo ha demostrado un alto nivel de precisión, con un F1 Score del 88.89%, recall del 84.60%, precisión del 95.89% y una accuracy del 95.69%.

El modelo está dirigido a una población específica en seguimiento por enfermedades precursoras como la diabetes mellitus tipo 2, hipertensión arterial y enfermedad renal crónica. Este modelo solo es aplicable a pacientes entre 18 y 80 años que estén siendo gestionados en una institución especializada y excluye a aquellos en terapia de reemplazo renal o con enfermedades autoinmunes.

OBJETIVO GENERAL: Validar externamente cuatro modelos predictivos diseñados para estimar daño renal a un año en población de pacientes crónicos colombianos.

DISEÑO DEL ESTUDIO Y FUENTES DE DATOS

Esta será un estudio retrospectivo longitudinal. Para la validación externa de los modelos se utilizarán datos de seguimiento de pacientes que fueron atendidos entre 2022 y 2024 en un prestador especializado en manejo de enfermedades crónicas. Estos datos comprenderán los mismos predictores usados en el desarrollo de los modelos, o en su defecto las mismas variables que permiten el cálculo de estos.

Se espera contar con una cohorte que permita incluir en los análisis datos de más de 1,000 pacientes, dado que un tamaño de muestra grande no solo permitirá realizar análisis más detallados y exhaustivos, sino que también permite una mayor seguridad en la generalización de los hallazgos a la población objetivo. Además, permitirá explorar y analizar subgrupos específicos con mayor detalle, lo que enriquecerá la comprensión de cómo los modelos se desempeñan en diferentes contextos clínicos o demográficos.

No obstante, el tamaño de muestra mínimo debería ser una muestra a conveniencia de 500 pacientes, con las mismas características de morbilidad de la cohorte incluida en el desarrollo y validación interna de los modelos, pero no necesariamente con las mismas características sociodemográficas. La validez externa de los modelos se asegurará al evaluar su rendimiento en una población diferente a la utilizada para desarrollarlos, lo que ayudaría a determinar su generalización y aplicabilidad en contextos clínicos diversos. Esto garantiza que el modelo sea válido y útil en entornos clínicos reales y para pacientes con diferentes perfiles demográficos.

El tamaño de muestra se estimó basado en la regla de un tamaño de muestra de al menos 100 eventos (casos positivos) en la muestra. Así, dada una prevalencia conocida de progresión acelerada de la ERC de entre el

20 y 40%, utilizando la fórmula para calcular el tamaño de muestra en estudios de proporciones $n=Z^2 \cdot p \cdot (1-p)/E^2$ (donde p=0.20 y Z=1.96 para un nivel de confianza del 95%, y E del 5% como margen de error deseado) el tamaño de muestra mínimo requerido sería de aproximadamente 500 pacientes para garantizar una buena estimación de los parámetros del modelo y evitar problemas de sobreajuste.

Al elegir la prevalencia más baja, se está adoptando un enfoque conservador que asegura que el tamaño de muestra sea lo suficientemente grande para capturar una muestra representativa de pacientes con ERC, sin subestimar el número necesario de casos positivos. Esta decisión se toma con el objetivo de minimizar el riesgo de errores de estimación y asegurar la robustez y la validez de los resultados del estudio. Además, al considerar la prevalencia más baja, se aumenta la probabilidad de obtener resultados que sean aplicables a una gama más amplia de pacientes con comorbilidades, lo que mejora la generalización de los hallazgos a la población objetivo.

Desenlaces y criterios de exclusion

Los desenlaces a predecir serán los mismos descritos en la descripción de los modelos validados internamente:

- Riesgo de progresión acelerada de la enfermedad renal en un lapso de 12 meses (categorizado como alto riesgo o bajo riesgo).
- Cambio de estadio renal en un periodo de 12 meses, categorizado en cinco niveles de riesgo (ERC1, ERC2, ERC3a, ERC3b, y ERC4).
- Tasa de filtración glomerular estimada (eTFG) en un horizonte de 12 meses.
- Rango de relación albuminuria creatinuria (RAC) en el que se ubicará cada paciente al cabo de un año.

La eTFG se estimará de dos formas, usando la ecuación de Cockcroft-Gault y la ecuación CKD-Epi. Se incluyen dos ecuaciones de modo que se pueda realizar un análisis de sensibilidad. Esto se realizará para evaluar cómo varía la predicción de la función renal cuando se emplean ecuaciones distintas en el cálculo. Dado que existen varias ecuaciones para estimar la eTFG y cada una tiene sus propias limitaciones y ventajas, realizar este análisis permitiría comprender mejor la robustez de los modelos predictivos ante diferentes métodos de estimación de la eTFG. Además se podrá determinar el impacto que la elección de la ecuación tiene en las predicciones de los modelos.

Los criterios de exclusion seran haber iniciado terapia de reemplazo renal, datos de seguimiento inferior a 12 meses, menos de dos mediciones de los predictores y el desenlace durante el seguimiento, diagnóstico de enfermedad renal poliquística, VIH, enfermedades autoinmunes, tratamiento contra el cáncer en los últimos 2 años y trasplante renal, según los registros médicos.

Predictores

- Sexo
- edad en años
- Peso en línea de base
- Peso actual

- Creatinina en suero en línea de base
- Creatinina en suero actual
- LDL actual
- HDL actual
- Triglicérido sérico actual
- Hemoglobina glicosilada actual
- Toma de medicamentos Antihipertensivos * (si / no)
- Toma de medicamentos Anti anémicos * (si / no)
- Toma de medicamentos Antidiabéticos * (si / no)
- Toma de medicamentos Antitrombóticos * (si / no)
- Toma de medicamentos Diuréticos * (si / no)
- Toma de medicamentos Modificadores Lípidos * (si / no)
- Diagnóstico de HTA (si / no)
- Diagnóstico de DM (si / no)
- Diagnóstico de Trastorno metabólico * (si / no)
- Diagnóstico de Trastorno Nutricional * (si / no)
- Diagnóstico de Trastorno Urológico * (si / no)
- eTFG en línea de base
- eTFG actual

Nota: *Variables construidas a partir de la agrupación de distintos registros en la historia clínica.

ANÁLISIS ESTADÍSTICO

Las estadísticas descriptivas para las variables predictoras se resumirán como media (± desviación estándar) o mediana (rango intercuartílico) para variables continuas o frecuencia (porcentaje) para variables categóricas. Las características de los participantes se compararan entre grupos utilizando la prueba de Chi-cuadrado para variables categóricas, y ANOVA de un factor o prueba de Kruskal-Wallis para variables continuas.

Tal y como se realizó en el desarrollo y validación interna de los modelos, los datos se divirán en dos partes: una para entrenamiento y otra para prueba. Se asignará aproximadamente el 70-80% de los datos para el conjunto de entrenamiento, y el 20-30% restante para el conjunto de prueba. Esta proporción permite que el modelo se entrene adecuadamente con una cantidad suficiente de datos, mientras que aún se reserva una porción significativa para evaluar su rendimiento en datos no vistos.

Se usarán las mismas tecnicas de aprendizaje automatico usadas en el desarrollo y validación interna, como bosques aleatorios, XGBoost y modelos de regresión de vectores de soporte (SVR).

Para los modelos que predicen un desenlace categorico, la discriminación se evaluará utilizando el estadístico c (Área bajo la curva ROC, AUC), que mide la capacidad para distinguir entre casos positivos y negativos. Las metricas de desempeño seran el F1 score, sensibilidad, especificidad, precision y exactitud que se compararán con las metricas obtenidas en la validación interna.

Para el modelo que predice un desenlace binario (progresión lenta y rápida de la ERC), se empleará el Puntaje Brier Escalado como métrica de evaluación de la calibración. Esta medida, derivada del Puntaje Brier original, es pertinente en contextos donde existe un desequilibrio en la distribución de clases (eventos

positivos son menos frecuentes) (Mogueo et al., 2015). El Puntaje Brier Escalado ajusta el puntaje Brier estándar para tener en cuenta la frecuencia de eventos positivos en los datos de prueba, proporcionando así una evaluación más precisa de la habilidad predictiva del modelo. Esta metrica permite evaluación integral y precisa de la capacidad del modelo para predecir el desenlace, considerando tanto la precisión como la calibración de las predicciones (Wu & Lee, 2014).

Para el modelo que predice un desenlace multiclase (cambio de estadio renal y RAC), se utilizarán el Brier Score Multiclase y el Gráfico de Reliabilidad Multiclase (Multiclass Reliability Plot) como métricas de evaluación de la calibración. El Brier Score Multiclase proporciona una medida de la discrepancia entre las probabilidades predichas por el modelo y las frecuencias observadas para todas las clases. Por otro lado, el Gráfico de Reliabilidad Multiclase permite visualizar cómo se comparan las probabilidades predichas con las frecuencias observadas para diferentes clases y rangos de probabilidad.

Para el modelo que predice un desenlace continuo (eTFG), se utilizaran el error absoluto medio (MAE), el error cuadrático medio (MSE), y el coeficiente de determinación R^2. Los valores que se obtengan tambien se compararán con los valores obtenidos en la validación interna.

Adicionalmente, para el modelo de eTFG se calculará error promedio de predicción absoluto (MAPE) (Tofallis, 2015). El MAPE mide las desviaciones entre los valores predichos y reales, siendo valores más pequeños indicativos de una mejor precisión en la predicción. MAPE se expresa típicamente como un porcentaje y proporciona una medida de la precisión relativa del modelo de predicción en comparación con los valores reales (Tofallis, 2015). Una de las ventajas del MAPE es que proporciona una medida de error en términos relativos, lo que lo hace útil para comparar la precisión de diferentes modelos o series de datos con diferentes escalas de valores.

Comparación por grupos de riesgo

Antes de la construcción de los modelos, con ayuda de un médico especialista se identificarán los subgrupos de interés, como diferentes grupos de edad, género, comorbilidades, etc.

Para conocer si el desempeño de los modelos difiere por grupos de riesgo (ej., hipertensos vs no hipertensos) se realizará un análisis exhaustivo del rendimiento de los modelos predictivos en subgrupos específicos de interés y de riesgo. Para lograr esto, se emplearán el análisis de calibración por subgrupos y el análisis de sensibilidad, especificidad, y curva ROC por subgrupos para el caso de los modelos que predicen un desenlace binario. Para aquellos modelos que predicen un desenlace multiclase se usará un análisis de precisión, recall y f1-score por subgrupos.

El análisis de calibración por subgrupos permitirá evaluar la calibración de los modelos en diferentes segmentos de la población, especialmente aquellos definidos por variables importantes como la presencia de comorbilidades y grupos de edad. este análisis ayudará a identificar si el modelo está bien calibrado en cada subgrupo específico, lo que es relevante para garantizar que las predicciones del modelo sean precisas y confiables en diversos contextos clínicos.

Por otro lado, el análisis de sensibilidad, especificidad y curva ROC por subgrupos permitirá evaluar cómo varía la capacidad del modelo para detectar verdaderos positivos y negativos en diferentes segmentos de la población. Este análisis proporcionará información sobre la sensibilidad y especificidad del modelo en cada subgrupo, así como la capacidad discriminativa global del modelo en diferentes contextos clínicos. El análisis de precisión, recall y F1-score por subgrupos para los modelos multiclase permitirá comprender

cómo se comporta el modelo en cada subgrupo específico en términos de su precisión, sensibilidad y precisión/sensibilidad.

El proceso de validación externa se llevará a cabo siguiendo la Lista de Verificación TRIPOD para la Validación de Modelos de Predicción (Collins et al., 2015). Esto es para garantizar la integridad y la robustez de la validación externa y para asegurar un informe claro y completo de los resultados obtenidos. Esta lista de verificación, reconocida internacionalmente, proporciona un marco detallado para garantizar la transparencia y la exhaustividad en la validación de modelos de predicción.

Análisis de sensibilidad

Se realizará una comparación de la tasa de filtración glomerular estimada (eTFG) calculada utilizando las ecuaciones de Cockcroft-Gault y CKD-Epi para cada paciente, empleando técnicas estadísticas como gráficos de dispersión, diagramas de Bland-Altman y análisis de sensibilidad para visualizar y evaluar las diferencias entre las estimaciones de eTFG obtenidas con cada ecuación.

En primer lugar, especificamente para el modelo predictivo de progresion acelerada de la ERC, se examinará cómo varía la predicción de la pérdida de función renal al emplear distintas ecuaciones, comparando dos modelos de regresion; uno que incluye como predictor la eTFG calculada usando una de las ecuaciones y otro modelo usando como predictor la eTFG calculada con la otra ecuación. Se evaluará el rendimiento de cada modelo, y se compararan los resultados en términos de las metricas de ajuste y visualizacion a través de gráficos de dispersión, diagramas de residuos, entre otros métodos.

En segundo lugar, para los demás modelos predictivos en donde la eTFG se incluye solo como predictor, se determinará el efecto de la elección de la ecuación en las predicciones de los modelos mediante análisis de regresión.

Análisis de sesgos y disparidad en el desempeño de los modelos:

- Identificación de subgrupos de riesgo.
- Evaluación del desempeño por subgrupos: Utilizando las métricas de desempeño del modelo (precisión, sensibilidad, especificidad, etc.), se evaluará cómo varía el desempeño del modelo en cada subgrupo.
- Interpretación de los resultados: Se analizarán los resultados para identificar posibles sesgos o disparidades en el desempeño del modelo entre los diferentes subgrupos.
- Discusión de las implicaciones: Se discutirán las implicaciones clínicas y prácticas de cualquier disparidad o sesgo detectado.

Validación de la interpretabilidad de los modelos:

Para validar la interpretabilidad de los modelos, se usara el gráfico SHapley Additive exPlanations (SHAP). Los valores SHAP proporcionan la contribución de cada predictor a cada predicción individual en un modelo propuesto y lo representa graficamente mostrando la importancia de cada predictor en los modelos.

En otras palabras, los valores SHAP cuantifican el impacto de cada predictor en las predicciones del modelo al considerar todas las posibles combinaciones de características.

- Evaluación por parte de expertos: Se solicitará la opinión de especialistas para evaluar la interpretabilidad del modelo y su capacidad para tomar decisiones clínicamente relevantes.
- Adaptación del modelo si es necesario: Si se identifican problemas de interpretabilidad, se realizarán ajustes en el modelo para mejorar su comprensión y confiabilidad, como la simplificación de características o la incorporación de explicaciones adicionales.

CONSIDERACIONES ÉTICAS

Antes de su inicio, el protocolo será sometido a una revisión exhaustiva por parte de un comité de ética en investigación para garantizar el cumplimiento de las normativas éticas y la protección de los datos. Además, se implementará la anonimización de todos los datos utilizados en los análisis, asegurando la confidencialidad y privacidad de la información de los pacientes involucrados en el estudio.

Por otro lado, se garantizará que los modelos predictivos sean transparentes y explicables, lo que significa que los procesos y criterios utilizados para desarrollarlos y validarlos se comunicaran de manera clara y comprensible para los usuarios finales, incluidos los profesionales de la salud y los pacientes. Esto es un paso crucial para generar confianza en los modelos y para facilitar su adopción y uso responsable.

Por último, con los análisis de las métricas de desempeño por subgrupos se pretende también evaluar sistemáticamente los modelos en términos de sesgo o disparidad en su desempeño entre diferentes grupos de pacientes. Si se detectan disparidades, se recomendarán medidas correctivas para abordarlas.

COMUNICACIÓN DE LOS RESULTADOS

En concordancia con la premisa de S4L de garantizar la transparencia, la confianza y la utilidad de los modelos predictivos desarrollados, se proponen las siguientes estrategias para comunicar los resultados tanto dentro de la organización como a otras partes interesadas relevantes:

1. Se preparará un informe detallado que resuma los hallazgos de la validación externa, incluyendo una descripción de la metodología utilizada, los resultados obtenidos, análisis de subgrupos y consideraciones éticas.
2. Se organizarán sesiones de presentación internas para compartir los resultados de la validación externa con los diferentes departamentos y equipos relevantes dentro de la organización. Estas presentaciones permitirán discutir los hallazgos de manera más interactiva, responder preguntas y recopilar retroalimentación de los miembros del equipo.
3. Publicación en Revistas Científicas: Se buscará la publicación de los resultados en al menos una revista científica especializada e indexada.
4. Presentaciones en congresos científicos: Se considerará la participación en conferencias y congresos relacionados con la salud pública, la epidemiología y IA en medicina, donde se puedan

presentar los resultados de la validación externa a audiencias especializadas y potencialmente interesadas en la implementación de los modelos.

Referencias

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Annals of Internal Medicine*, *162*(1), 55–63. https://doi.org/10.7326/M14-0697

Mogueo, A., Echouffo-Tcheugui, J. B., Matsha, T. E., Erasmus, R. T., & Kengne, A. P. (2015). Validation of two prediction models of undiagnosed chronic kidney disease in mixed-ancestry South Africans. *BMC Nephrology*, *16*(1), 94. https://doi.org/10.1186/s12882-015-0093-6

Tofallis, C. (2015). A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society*, *66*(8), 1352–1362. https://doi.org/10.1057/jors.2014.103

Wu, Y.-C., & Lee, W.-C. (2014). Alternative Performance Measures for Prediction Models. *PLoS ONE*, *9*(3), e91249. https://doi.org/10.1371/journal.pone.0091249