UNIVERSIDAD
NACIONAL
DE COLOMBIA

Detección temprana de cáncer en población sintomática: desarrollo de un modelo multivariante y un modelo de aprendizaje de máquinas basados en bases de datos administrativas del sistema de salud colombiano y aplicados al cáncer gástrico

Raúl Hernando Murillo Moreno; MD, MSP

Facultad de Medicina
Departamento de Patología

Bogotá DC; 2024

Detección temprana de cáncer en población sintomática: desarrollo de un modelo multivariante y un modelo de aprendizaje de máquinas basados en bases de datos administrativas del sistema de salud colombiano y aplicados al cáncer gástrico

(*Cancer early detection in symptomatic patients: development of multivariate and machine learning models based on administrative health data of the Colombian health system and applied to gastric cancer)*

Tesis presentada para optar al título de
Doctor en Oncología

Presentado por

Raúl Hernando Murillo Moreno

Director de tesis

Giancarlo Buitrago

Co-director de tesis
Diego Patiño

Grupos de Investigación

Grupo de Oncología – PUJ-HUSI

CEPIT – Pontificia Universidad Javeriana

Grupo de Investigaciones Clínicas en Cáncer - Universidad Nacional de Colombia

Bogotá DC, mayo 2024

Dedicatoria

A mi madre María Felisa,
a quien me debo
a mi esposa Luz Helena,
y a mi hijo Nicolás,
quienes me han regalado de su tiempo
y me han cubierto con su afecto
durante mi tránsito por este sendero

Tabla de contenido

Summary i
1. Resumen del proyecto iv
2. Planteamiento del problema 1
2.1 Descripción de la realidad del problema 1
2.2 Formulación del problema 3
3. Justificación 4
4. Objetivos 5
5. Marco teórico 5
5.1 Historia natural del cáncer gástrico 5
5.2 Detección temprana del cáncer gástrico 7
5.3 Modelos predictivos en detección temprana de cáncer 8
5.3.1 Escalas de riesgo a partir de factores de riesgo establecidos 9
5.3.2 Escalas de riesgo a partir de síntomas 10
5.4 Consideraciones metodológicas 12
5.4.1 Uso de bases de datos administrativas 12
5.4.2 Modelos de predicción clínica 13
5.4.3 Aprendizaje de máquinas 15
6. Hipótesis 17
7. Diseño metodológico 17
7.1 Diseño general del estudio 17
7.2 Fuentes de información 18
7.3 Definición de sujetos de estudio 20
7.4 Limpieza de bases de datos y selección de periodos de análisis 20
7.5 Etiquetado de casos y controles 22
7.5.1 Análisis exploratorio del comportamiento de las etiquetas 22
7.5.2 Determinación de sensibilidad y especificidad de la codificación CIE-10 24
7.5.3 Validación de etiquetas mediante estimación de prevalencias 28
7.6 Construcción de un modelo multivariado de predicción clínica 29
7.6.1 Análisis exploratorio para la identificación de variables predictoras 30
7.6.2 Etiquetado de sujetos a riesgo y controles 32
7.6.3 Selección de la muestra y fusión de bases de datos 34
7.6.4 Análisis de asociación y riesgos competitivos 36

7.7 Construcción de un modelo predictivo basado en aprendizaje de máquinas 37
7.7.1 Selección de pacientes y estructuración de la base de datos 37
7.7.2 Etiquetado de casos y controles 38
7.7.3 Pre-procesamiento de datos (filtros para selección de variables) 38
7.7.4 Procesamiento: desarrollo de los modelos de aprendizaje de máquinas 40
7.7.5 Evaluación del os modelos predictivos 43
8. Resultados 43
8.1 Modelo multivariante 43
8.2 Modelo de aprendizaje de máquinas 46
9. Discusión 52
Implicaciones para la práctica clínica 60
Limitaciones del estudio 61
10. Conclusiones 64
11. Agradecimiento 65
12. Anexos 67
13. Bibliografía 80

Lista de tablas

Tabla 1. Características de algoritmos diagnósticos desarrollados para la detección de cáncer en cuidado primario 11

Tabla 2. Disponibilidad de datos discriminada por año en la base de datos 20

Tabla 3. Registros totales y registros faltantes para fecha de nacimiento y sexo discriminados por año en las bases de datos disponibles 21

Tabla 4. Número de registros según tipo de etiquetado para el año 2013 24

Tabla 5. Disponibilidad y utilización de variables para el emparejamiento de registros 25

Tabla 6. Emparejamiento total y parcial de registros 26

Tabla 7. Comparación de registros emparejados y no emparejados 27

Tabla 8. Estimación de casos de cáncer gástrico prevalentes en Bucaramanga 29

Tabla 9. Distribución por edad y sexo de la muestra para el análisis exploratorio de variables predictoras en la construcción del modelo multivariante 31

Tabla 10. Número de variables por categoría que mostraron asociación significativa con el diagnóstico de cáncer gástrico en el análisis exploratorio para la construcción del modelo multivariante 31

Tabla 11. Diagnósticos asociados positivamente al diagnóstico de cáncer gástrico y procedimientos y medicamentos relacionados con los diagnósticos en el análisis exploratorio para la construcción del modelo multivariante 33

Tabla 12. Estructura de las bases de datos para la construcción del modelo multivariante 35

Tabla 13. Comorbilidad considerada en el análisis del modelo multivariante y codificación para construcción del índice de comorbilidad de Charlson 36

Tabla 14. Agrupación de códigos diagnósticos (CIE10) para la construcción del modelo basado en aprendizaje de máquinas 40

Tabla 15. Características de la población de estudio para construcción del modelo multivariante 43

Tabla 16. Medidas de asociación (valores no ajustados) del antecedente de anemia con el diagnóstico de cáncer gástrico 44

Tabla 17. Medidas de asociación ajustadas para las variables incluidas en el modelo 45

Tabla 18. Razón de riesgos de cáncer gástrico para las variables en el modelo de riesgos competitivos según Fine y Gray 46

Tabla 19. Características de la población de estudio en los modelos de aprendizaje de máquinas según periodo de análisis 47

Tabla 20. Desempeño de los modelos de aprendizaje de máquinas en el estudio .. 49

Tabla 21. Síntomas y diagnósticos reportados como variables predictivas en modelos de aprendizaje de máquinas seleccionados 57

Lista de gráficos

Gráfico 1. Tendencia de la cobertura de aseguramiento según régimen de seguros durante el periodo de estudio en Colombia 18

Gráfico 2. Distribución de la edad por periodo anual en las bases de datos disponibles 21

Gráfico 3. Tendencia del uso de servicios (atenciones) en la población de estudio durante el periodo de observación 44

Gráfico 4. Incidencia acumulada de cáncer gástrico en sujetos con y sin anemia. Cálculo según modelo de riesgos competitivos 45

Gráfico 5. Distribución de los coeficientes para las variables incluidas en el modelo (diagnósticos y procedimientos) 48

Gráfico 6. Curva ROC de los modelos de aprendizaje de máquinas en el estudio ... 50

Gráfico 7. Contribución de las variables con los mayores y menores coeficientes a la capacidad predictiva de los modelos de aprendizaje de máquinas 51

Gráfico 8. Relación entre población a seguir y porcentaje de casos de cáncer gástrico detectados según los modelos de aprendizaje de máquinas en el estudio 52

Lista de figuras

Figura 1: Historia natural de la infección por H pylori 6

Figura 2. Flujograma para la construcción del modelo multivariante 30

Figura 3. Flujograma para la construcción del modelo de aprendizaje de máquinas 37

Figura 4. Representación esquemática de los modelos de aprendizaje de máquinas utilizados en el estudio 40

Figura 5. Flujograma del proceso de selección de la muestra en aprendizaje de máquinas 46

Summary

Stomach cancer is the leading cause of cancer death in Colombia. There is no evidence-based screening available for this condition and early diagnosis in symptomatic patients is a more suitable alternative for low- and middle-income countries given the lower burden for health systems and patients.

Predictive models have been developed but restricted to variables with very high positive predictive value. Artificial intelligence makes possible the search for new alternatives for stomach cancer early detection; however, most studies on this topic focus on endoscopic images or improvement of pre-existing epidemiological models. In addition, we found no studies from Latin America.

Using databases from the Colombian health system, this study analyzes the association of stomach cancer diagnosis with previous clinical diagnoses and procedures, based on the idea that the systematic recording of all clinical conditions allows us to identify the patient's footprint during the carcinogenic process. A study with two components was developed: a multivariate clinical prediction model (prospective) and a prediction model based on machine learning (retrospective).

We used the database for the definition of the Capitation Payment Unit in the contributory regime of the Colombian health system (BDUPC given its Spanish acronym) and the country's death information from the vital statistics system. The period of analysis was defined based on data availability completeness, comparability (coding), and consistency.

Multivariate model (prospective analysis): Subjects aged 40 to 90 years old who attended health services in 2011 were included. Patients with cancer diagnosis and patients reported in the mortality database for the year 2011 were excluded.

LASSO logistic regression was used to select predictive variables. Non-hemolytic anemia was defined as the exposure variable and the Charlson comorbidity index a covariate (17 diagnoses). In total, 14,399 individuals with anemia and 44,961 without anemia made up the follow-up cohort. The risk of stomach cancer diagnosis was estimated using a cumulative incidence function, and subsequently a competing risks model.

The groups showed differences in age, sex and place of residence. The average number of visits to health services was 1.8; 23.0 and 40.3 for patients without anemia, with anemia and with gastric cancer, respectively. The relative risk for stomach cancer diagnosis was 1.53 (95% CI 1.51-1.54), with attributable fraction 47.1 and a population attributable fraction 17.8. The median time to diagnosis for subjects with anemia was 28.2 months (IQR 17.2-40.9) and for subjects without anemia 36.7 months (IQR 23.8-48.0). The cumulative incidence of gastric cancer at 60 months in subjects with anemia was 7.6 per 1,000 (95% CI 6.1-9.1) and in subjects without anemia 2.4 per 1,000 (95% CI 1.9-2 ,8). In the competing risks model, anemia showed the greatest strength of association followed by male sex.

Machine Learning Model (retrospective Analysis): A sample of patients with and without gastric cancer attending health services between September 1st and December 31st/2017 was selected. We included patients 30 to 90 years old. We analyzed the predictive capacity of the models close to the outcome (January-August 2017), and in a two-year period prior to the outcome (September 2015-August 2017). In the latter case,

a washout period of six months prior to the outcome was allowed in order to reduce the effect of diagnoses and procedures close to the outcome.

Predictive variables (diagnoses by CIE10 codes and procedures by CUPS codes) were filtered using LASSO logistic regression. In total 1,426 and 1,462 variables were included for the one-year and two-year periods of analysis, respectively. Three machine learning models were used: LASSO logistic regression, random forests, and gradient boosting. The accuracy of the models was evaluated by determining the area under the curve (AUROC), accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Additionally, the number needed to follow-up was calculated (NNT).

After filtering by data quality, 4,965,931 and 5,175,620 individuals were included for the one-year and two-year periods of analysis, respectively. There were 1,426 and 1,462 cases of stomach cancer, respectively. Most patients were women, the average age was 58 years in patients with gastric cancer and 53 years in controls. The rate of health service attendance was higher in patients with gastric.

The coefficients for diagnoses and procedures were higher for the one-year period. Diagnostic codes showed greater predictive capacity than procedures codes. The machine learning models showed no difference in AUROC; however, when using the best cutoff point for each model, the sensitivity for LASSO was higher than in random forests and gradient boosting (65.4%; 27.8% and 57. 9%, respectively). Specificity was higher in the random forest at the cost of a significant reduction in sensitivity.

The predictive capacity of the LASSO model for the one-year period was the highest. In all cases, the accuracy improved with a higher cut-off point (greater probability of predicting the outcome), essentially due to an increase in specificity. Due to the low prevalence of the disease (approximately 0.02%), the PPV in all models was lower than 1%. The AUROC did not show relevant differences between training and test datasets.

In all models, municipalities located in the Colombian mountains reported higher coefficients, and municipalities located in the Atlantic Coast lower coefficients.

The detect at least 70% of stomach cancer cases, we need to follow-up 16.1%; 16.5%; and 20.4% of the population during two years according to LASSO, random forests and gradient boosting, respectively. Such numbers are equivalent to a NNT 1,170, 1,230, and 1,498, respectively. The LASSO logistic regression at one-year reports that 10.6% of the population requires follow-up to detect at least 70% of stomach cancer cases (NNT 790).

Conclusions: The machine learning models developed in this study represent an innovative approach by using all available data related to the symptomatic expression of the disease in the pre-diagnostic phase. LASSO showed better performance than other models (including the multivariate model). The observed performance might improve in a higher risk population such as residents in high mountain municipalities. The results have potential for use in clinical practice given their discriminative capacity and the NNT, but their use in routine services must be preceded by their correct clinical validation.

Key words: stomach neoplasms, early detection of cancer, machine learning, big data, routinely collected health data, Colombia

1. Resumen del proyecto

En Colombia, el cáncer gástrico es la primera causa de muerte por cáncer. En el momento no existen estrategias de tamización con eficacia demostrada para esta enfermedad y mejorar la oportunidad diagnóstica en sintomáticos representaría menor esfuerzo para sistemas de salud y pacientes.

Diversos modelos utilizan síntomas y factores de riesgo restringidos a variables con muy alto valor predictivo positivo. La inteligencia artificial posibilita la búsqueda de nuevas alternativas, pero los estudios en este campo se orientan a mejorar la interpretación de imágenes endoscópicas o a mejorar modelos epidemiológicos preexistentes. Adicionalmente, no se conocen trabajos desarrollados en Latinoamérica.

Utilizando bases de datos del sistema de salud, este trabajo abordó la problemática descrita mediante el análisis de la asociación del diagnóstico de cáncer gástrico con diagnósticos y procedimientos registrados en todas las atenciones previas al diagnóstico de cáncer, sobre la idea de que el registro sistemático de todas las atenciones permite identificar el rastro que deja el paciente durante el desarrollo de la patología oncológica. Se realizó un estudio con dos componentes: modelo multivariado de predicción clínica (prospectivo) y modelos de predicción basados en aprendizaje de máquinas (retrospectivo).

Se utilizó la base de datos para cálculo de la Unidad de Pago por Capitación en el régimen contributivo del sistema de salud y la información de defunciones del sistema de estadísticas vitales. Se seleccionaron periodos de análisis con base en la disponibilidad, suficiencia, comparabilidad y consistencia de los datos.

Modelo multivariante: Se incluyeron pacientes de 40 a 90 años que usaron servicios en 2011. Se excluyeron pacientes con diagnóstico de cáncer o fallecidos ese año.

Se utilizó regresión logística LASSO para seleccionar variables predictoras. Las anemias carenciales se definieron como variable de exposición, y el índice de comorbilidad de Charlson modificado como covariable. En total conformaron la cohorte 14.399 individuos con anemia (expuestos) y 44.961 sin anemia (no expuestos). Se analizó el riesgo de diagnóstico de cáncer gástrico mediante una función de incidencia acumulada. Posteriormente se utilizó un modelo de riesgos competitivos.

Se observaron diferencias en edad, sexo y lugar de residencia entre expuestos y no expuestos. El riesgo relativo para el diagnóstico de cáncer gástrico fue 1,53 (IC95% 1,51-1,54), con fracción atribuible 47,1 y fracción atribuible poblacional 17,8. La mediana de tiempo al diagnóstico fue 28,2 y 36,7 meses para expuestos y no expuestos. La incidencia de cáncer gástrico a 60 meses con anemia fue 7,6 por 1.000 (IC95% 6,1-9,1) y sin anemia 2,4 por 1.000 (IC95% 1,9-2,8). En el modelo de riesgos competitivos la anemia mostró la mayor fuerza de asociación.

Modelo de aprendizaje de máquinas: Se seleccionaron pacientes de 30 a 90 años con y sin diagnóstico de cáncer gástrico entre septiembre y diciembre de 2017. Se utilizaron dos periodos de análisis: próximo al desenlace (enero-agosto 2017) y dos años previos al desenlace (septiembre 2015-agosto 2017). En el último caso se dejó un periodo de lavado de seis meses para reducir el efecto de diagnósticos y procedimientos cercanos al desenlace.

Se filtraron variables predictoras (CIE10 diagnósticos y CUPS procedimientos). Se incluyeron 1.426 y 1.462 variables para los análisis a uno y dos años, respectivamente. Se utilizaron tres modelos: regresión logística LASSO, bosques aleatorios y potenciación del gradiente. Se evaluó el desempeño mediante el área bajo la curva (AUROC), exactitud (accuracy), sensibilidad, especificidad, valor predictivo positivo (VPP) y valor predictivo negativo (VPN). Se calculó el número necesario a seguir (NNT) en función de la probabilidad de diagnóstico de cáncer.

Se incluyeron 4.965.931 y 5.175.620 individuos para el seguimiento a uno y dos años, dentro de los cuales hubo 1.426 y 1.462 casos de cáncer gástrico, respectivamente. La mayoría fueron mujeres, la edad promedio 58 y 53 años en pacientes con y sin cáncer gástrico, el número de atenciones fue mayor en pacientes con cáncer gástrico.

Los coeficientes para las variables predictivas fueron mayores en el análisis de un año. A su vez, los diagnósticos mostraron mayor capacidad predictiva en ambos periodos de análisis. El AUROC no mostró diferencias entre modelos en el análisis a dos años; sin embargo, al utilizar el mejor punto de corte de cada modelo, la sensibilidad para LASSO fue mayor que para bosques aleatorios y potenciación de gradiente (65,4%; 27,8% y 57,9%). La especificidad fue mayor en los bosques aleatorios.

La mayor capacidad predictiva la tuvo la regresión logística LASSO en el año previo al diagnóstico. En todos los modelos, el desempeño se mejoró con el incremento del punto de corte (mayor probabilidad de predicción del desenlace), esencialmente por un incremento de la especificidad. Los VPP en todos los modelos fueron menores de 1% en razón a la baja prevalencia de la enfermedad. El AUROC no mostró diferencias entre el ejercicio de validación interno (entrenamiento) y la validación externa.

Consistentemente, municipios ubicados en la zona de montaña estuvieron dentro de las variables con coeficientes más altos mientras que municipios de la Costa Atlántica estuvieron dentro de las variables con coeficientes más bajos.

En el análisis a dos años, la proporción de individuos que requieren seguimiento para detectar al menos el 70% de los casos de cáncer gástrico es 16,1%; 16,5%; y 20,4% según LASSO, bosques aleatorios y potenciación del gradiente, respectivamente; esto equivale a un NNT de 1.170, 1.230, y 1.498. La regresión logística LASSO a un año muestra que se debe seguir el 10,6% de la población para detectar al menos el 70% de los casos de cáncer gástrico (NNT 790).

Conclusiones: El presente estudio representa un abordaje innovador del aprendizaje de máquinas al utilizar información no sesgada de toda expresión sintomática de la enfermedad en su fase previa al diagnóstico. La regresión logística LASSO mostró mejor desempeño que los modelos restantes incluido el modelo multivariante. El desempeño observado puede mejorarse en población de mayor riesgo como los residentes en los municipios de alta montaña. Los resultados tienen potencial de uso en la práctica clínica dada su capacidad discriminativa y el NNT, pero su uso en la rutina de los servicios debe estar precedido de su correcta validación clínica.

Palabras clave: neoplasias gástricas, detección precoz del cáncer, aprendizaje de máquinas, datos masivos, datos de salud recolectados rutinariamente, Colombia

2. Problema de investigación

2.1 Descripción de la realidad del problema

El cáncer representa actualmente una de las mayores causas de enfermedad y muerte en el mundo. En Colombia, cada año se observan alrededor de 113.000 casos nuevos y 55.000 muertes por esta causa y se espera para el año 2040 un incremento en el número de casos y muertes solo por efecto del cambio demográfico(1).

En este escenario, el cáncer gástrico es la tercera causa de muerte por cáncer en los países de ingresos medios y bajos, y continúa siendo la primera causa de muerte por cáncer en Colombia (12,2% del total de muertes por cáncer)(2). Aun asumiendo la tendencia decreciente en la última década pasaremos de tener cerca de 8220 casos nuevos por año en 2020 a 13800 en 2040, y de 6450 a 10860 muertes respectivamente (CPPA incidencia -0,08 en hombres y -0,83 en mujeres, CPPA mortalidad -0,8 en hombres y -0,12 en mujeres)(2).

Si se mantienen esta tendencia no tendríamos cambio en la razón mortalidad/incidencia a 2040 (0,77 en 2020 y 0,79 en 2040), lo que indica que se mantendría la elevada letalidad de la enfermedad en los próximos 20 años, letalidad que refleja la baja proporción de casos diagnosticados en estados tempranos (23,8% en 2022)(3) y se expresa en una baja supervivencia a 3 y 5 años (36,8% y 17,1% respectivamente)(4).

A pesar de que la supervivencia es mayor en países europeos, los valores observados no superan el 35% y la razón mortalidad/incidencia es relativamente alta, lo que sugiere una elevada letalidad de la enfermedad aún en escenarios con mayores recursos y mejor desempeño de los sistemas de salud. La detección temprana sigue entonces siendo pilar fundamental para el control de la enfermedad y constituye base esencial de la racionalidad en el uso de recursos. Diversos estudios muestran cómo el costo al final de la vida para los pacientes con cáncer es mayor que en etapas tempranas de la enfermedad (5), e invariablemente, la supervivencia y las tasas de curación son mayores para los estados tempranos de la enfermedad (supervivencia neta a 3 años 58,1% y 25,1% para estados I-II y estado IV respectivamente)(6), aún con el advenimiento de las terapias dirigidas a blancos moleculares.

No existen en el momento estrategias de tamización con eficacia demostrada sobre la reducción de mortalidad (7). Los únicos países que han implementado tamizaje masivo son Japón y Corea del Sur, pero en ninguno de los dos casos la implementación estuvo precedida de estudios clínicos, y a pesar de la baja relación mortalidad/incidencia (0,26 y 0,22 respectivamente)(8), los datos de vida real no logran distinguir el efecto de la tamización del efecto de los cambios en los determinantes sociales sobre la reducción de la mortalidad (7), de forma similar a lo ocurrido en Colombia, en donde se observa una reducción en la mortalidad sin intervenciones para la detección temprana y sin modificación de la letalidad (8).

Adicional a las limitaciones de la evidencia disponible, la tamización de cáncer gástrico implicaría la realización repetida de exámenes rutinarios a una franja inmensa de individuos asintomáticos, para lo cual resulta indispensable una elevada inversión de recursos por parte de los sistemas de salud y un elevado nivel de organización que posibilite la identificación y seguimiento de cada individuo en la población objeto, el

llamado, el re-llamado ante resultados inconclusos y el seguimiento estricto de los resultados positivos al tamizaje (9). Tal esfuerzo limita de forma importante la posibilidad de implementar programas de tamización efectivos en la mayoría de países de ingresos bajos y medios, en donde el nivel de organización para la tamización de otros tipos de cáncer sigue siendo deficiente, resultando en una inversión considerable de recursos sin mayor impacto.

Una alternativa a esta situación es la posibilidad de mejorar la oportunidad del diagnóstico en personas sintomáticas, lo que representa un esfuerzo menor tanto para los sistemas de salud como para los individuos, y podría por ende generar una relación costo-beneficio más favorable. De hecho, la mayor supervivencia observada en países de ingresos altos se debe esencialmente a una mayor oportunidad en el diagnóstico de pacientes sintomáticos y en el acceso a tratamiento adecuado (4).

En cáncer gástrico, el proceso carcinogénico de la infección por *H pylori* ha sido claramente establecido. A pesar de que la progresión hacia cáncer puede cursar sin síntomas, la infección tiene múltiples desenlaces clínicos que generan un amplio espectro sintomático, los que incluyen úlcera péptica, gastritis, dispepsia (10), y estados patológicos intermedios en el proceso de carcinogénesis como la atrofia gástrica y las lesiones precursoras y preneoplásicas como la metaplasia y la displasia (11).

Esta circunstancia plantea un reto mayor al diagnóstico temprano de la enfermedad en población sintomática. Ante ello, se han propuesto modelos predictivos del riesgo de desarrollar la enfermedad, a fin de hacer una mejor selección de las personas que deben ir a tamización. Estos modelos han incorporado diferentes variables predictoras incluyendo características sociodemográficas, factores de riesgo, y cambios en la mucosa gástrica (12, 13). Adicionalmente, diversos estudios han intentado desarrollar modelos predictivos a partir de la presencia conjunta de síntomas y factores de riesgo (14). Algunas limitaciones en la generación de conocimiento desde esta perspectiva incluyen el bajo número de observaciones (desenlace) que requiere grandes tamaños muestrales, y la selección restringida y a priori de los síntomas a investigar desde la perspectiva del investigador, los que además se analizan usualmente de forma individual.

El auge de la inteligencia artificial ha promovido la búsqueda de alternativas para la detección temprana de la enfermedad. La mayoría de estudios se han hecho en torno al mejoramiento de la interpretación de imágenes endoscópicas como prueba de tamización y diagnóstico (15, 16). Adicionalmente, se han desarrollado modelos predictivos para estimar el riesgo de cáncer gástrico usando aprendizaje de máquinas, a fin de identificar sujetos que deberían entrar en un protocolo de seguimiento clínico (17-20); a la fecha estos modelos han utilizado información sobre un número limitado de variables incluyendo datos sociodemográficos, factores de riesgo establecidos para cáncer gástrico, síntomas del espectro clínico conocido para la enfermedad, pruebas de laboratorio para condiciones asociadas, y antecedentes patológicos seleccionados. La comparación con modelos epidemiológicos no muestra consistentemente un mejor desempeño para el aprendizaje de máquinas (21); sin embargo, este último se ha basado igualmente en la selección a priori de las variables a incluir como en los modelos epidemiológicos tradicionales y en su mayoría utilizan bases de datos prediseñadas para el estudio de cáncer gástrico. Adicionalmente, la gran mayoría se ha desarrollado en poblaciones asiáticas y no hay ninguno conocido en población latinoamericana (17), lo

cual representa una limitante mayor para su potencial implementación dada la necesidad de validación que requiere bases de datos similares a las de los estudios originales, algo poco probable de encontrar en países de ingresos medios y bajos.

Actualmente, la existencia de grandes bases datos poblaciones con información recolectada durante largos periodos de tiempo, abre puertas a nuevos escenarios de investigación. La disponibilidad de esta información reduce el sesgo de la selección a priori de variables para la inclusión en los modelos predictivos, y reúne gran cantidad de observaciones que elimina los problemas de tamaños muestrales. En este contexto, este trabajo de investigación abordó la problemática descrita utilizando las bases de datos del sistema de salud colombiano, mediante el análisis de la asociación del diagnóstico de cáncer gástrico con los motivos de consulta expresados en diagnósticos clínicos asociados a las atenciones previas recibidas, sobre la idea de que el registro sistemático de cualquier condición clínica permite identificar el rastro que ha dejado el paciente en el sistema de salud durante el proceso de desarrollo de la patología oncológica.

2.2 Formulación del problema

A pesar de constituir un pilar fundamental en el control de la enfermedad, la detección temprana de cáncer gástrico sigue representando un reto para los sistemas de salud por:

1. La inexistencia de estrategias de tamización con eficacia demostrada
2. La elevada inversión de recursos y el alto nivel organizativo que requeriría la tamización de base poblacional si se contara con una estrategia probada
3. El bajo nivel predictivo de los síntomas reconocidos dentro del espectro clínico de la enfermedad, los que además tienen mayor prevalencia en estados avanzados de la enfermedad.
4. Las limitaciones de los abordajes epidemiológicos tradicionales específicamente referidas al sesgo inducido por la selección a priori de variables y el requerimiento de grandes tamaños muestrales que reducen la factibilidad de estudios longitudinales.

Ante esta situación se plantearon las siguientes preguntas de investigación:

1. ¿Es posible diagnosticar el cáncer gástrico en etapas más tempranas a partir de la expresión de síntomas que motivan la consulta a los servicios de salud durante el desarrollo de la enfermedad?

Esta pregunta se aborda en el presente trabajo a partir de la búsqueda de respuestas a tres preguntas iniciales en el proceso de investigación

2. ¿Existe asociación entre motivos de consulta expresados en diagnósticos clínicos y el diagnóstico posterior de cáncer gástrico?
3. ¿Existe asociación entre motivos de consulta expresados en el uso de servicios de salud específicos y el diagnóstico posterior de cáncer gástrico?
4. ¿Cuál es el desempeño de modelos predictivos construidos a partir de la asociación identificada entre diagnósticos y servicios utilizados previamente al diagnóstico de cáncer gástrico?

3. Justificación

Como se expresó, Colombia experimentará un incremento en el número absoluto de casos y de muertes por cáncer gástrico debido al envejecimiento poblacional. Los datos disponibles sugieren que la reducción en la mortalidad observada en el país se presenta asociada a la reducción en la incidencia, indicando que esta última es su mayor determinante (22). El cáncer gástrico no ha tenido ninguna intervención preventiva desde los servicios de salud que explique la reducción en la incidencia, por lo que podría asumirse que dicha tendencia se explica principalmente por cambios en los determinantes socio-económicos.

El mínimo impacto de los servicios de salud en el control de la enfermedad coincide el deterioro de la supervivencia (4, 6). A diferencia de los observado en Colombia, en la mayoría de países de ingresos altos la reducción de la mortalidad ha sido sustancialmente mayor que la reducción de la incidencia, lo que se traduce en una menor razón mortalidad/incidencia e indica un efecto considerable de los servicios de salud(22). Tal efecto puede atribuirse a una mayor oportunidad en el diagnóstico, ya que, a diferencia de otros tipos de cáncer, en cáncer gástrico no ha habido un incremento significativo de la supervivencia asociado a las nuevas alternativas terapéuticas (4)

La detección temprana mediante tamización o mediante la búsqueda de la enfermedad en pacientes sintomáticos requiere un elevado nivel de organización y una alta capacidad de respuesta de los sistemas de salud. Adicionalmente, el bajo valor predictivo de los síntomas individuales hace que el diagnóstico temprano de tumores agresivos como el de estómago siga siendo un reto aún para los sistemas de salud con elevado nivel de organización. En consecuencia, el desarrollo de algoritmos diagnósticos con mayor valor predictivo que la expresión de síntomas individuales, ayudaría de forma significativa a la identificación de pacientes que deben ser objeto de un escrutinio diagnóstico o de un seguimiento más intensivo.

De lograrse el desarrollo de tales algoritmos, el desempeño de los sistemas de salud frente la detección temprana del cáncer gástrico tendría una mejora significativa, toda vez que el esfuerzo y los recursos dedicados al seguimiento y examen de pacientes se reduce de forma sustancial frente a lo requerido en los programas de tamización, y al mismo tiempo posibilita una mejor orientación para la implementación de rutas de atención que dependan en menor grado del nivel de organización y desarrollo de los sistemas de salud por ser más expeditas y brindar más claridad frente al riesgo de cáncer gástrico para los actores involucrados en el proceso diagnóstico en diferentes niveles del sistema de salud. En tal sentido, un efecto potencial del desarrollo de tales algoritmos es el fortalecimiento de los niveles primario y secundario de atención, los cuales, dada la problemática descrita juegan hoy un papel secundario en el diagnóstico temprano de la enfermedad por fuera de la tamización, y se ven de hecho como uno de los factores contribuyentes al diagnóstico tardío de la misma.

Desde otra perspectiva, la mejora en el desempeño de pruebas diagnósticas mediante su combinación con factores de riesgo, o la selección de individuos asintomáticos de alto riesgo para centrar la tamización en esta población, tienen limitaciones derivadas de la baja fuerza de asociación de los factores de riesgo individuales y de la fracción atribuible poblacional por la que la mayoría de los casos provienen de la población de bajo riesgo

en lo que se conoce como la paradoja preventiva (23). A diferencia de dichos abordajes, el presente trabajo de investigación no se basó en el riesgo de desarrollar la enfermedad sino en la expresión sintomática temprana de la misma. El abordaje propuesto se basó en el registro de todas las atenciones recibidas por los afiliados al sistema de salud colombiano (régimen contributivo), para la búsqueda de perfiles de enfermedad a partir del tipo de servicios utilizados y de los diagnósticos asociados a los mismos. La propuesta se basó en la probabilidad de que las manifestaciones clínicas tempranas del cáncer gástrico no se hayan asociado debido a la distancia temporal frente a la confirmación diagnóstica o sencillamente por la inexistencia de información, pero que de cualquier forma dejan registro de la trayectoria de atenciones recibidas por el paciente durante el desarrollo de la neoplasia. La construcción de algoritmos diagnósticos desde esta perspectiva, posibilita una aproximación diferente en la investigación y desarrollo de pruebas para la detección temprana del cáncer.

4. Objetivos

4.1 Objetivo primario

Establecer la capacidad predictiva para el diagnóstico de cáncer de estómago de algoritmos basados en las trayectorias de atención y sus diagnósticos asociados en pacientes que consultan al sistema de salud, mediante la construcción de un modelo multivariado.

4.2 Objetivos secundarios

- Evaluar la capacidad predictiva de los algoritmos diagnósticos construidos a partir de la trayectoria de atención en el sistema de salud utilizando modelos multivariados, en una condición sindromática diferente de la del cáncer de estómago y caracterizada por baja expresión de síntomas tempranos como es el cáncer de páncreas.
- Contrastar la capacidad predictiva de los algoritmos diagnósticos construidos con base en la trayectoria de atención en el sistema de salud mediante métodos epidemiológicos, con un modelo que utiliza la misma información, pero se basa en los métodos del aprendizaje de máquinas.

5. Marco teórico

5.1 Historia natural del cáncer gástrico

Los dos principales tipos histológicos del cáncer de estómago son el intestinal y el difuso, los que se han asociado con procesos etiopatogénicos diferentes y por ende con comportamiento epidemiológico diferente (24). El cáncer de tipo difuso tiene mayor participación porcentual en zonas de bajo riesgo poblacional y tiene mayor asociación con cáncer heredofamiliar. El cáncer de tipo intestinal se asocia más con el cáncer esporádico o de origen ambiental y tiene mayor asociación con la infección por *H pylori*, correspondiendo en Colombia al 74,9% de los casos (25).

El proceso carcinogénico de la infección por *H pylori* ha sido claramente establecido (Figura 1). El *H pylori* es una bacteria gram-negativa con elevada prevalencia de infección a nivel mundial, la que varía entre 20% y 50% en países de altos ingresos y se

encuentra con frecuencia por encima de 80% en países de ingresos medios y bajos (26). En población colombiana se reportó una prevalencia cercana a 70% en 2003 (27); sin embargo, la prevalencia se ha reducido globalmente debido a factores como el mejoramiento de las condiciones socioeconómicas y sanitarias, así como al uso extendido de antibióticos (10).

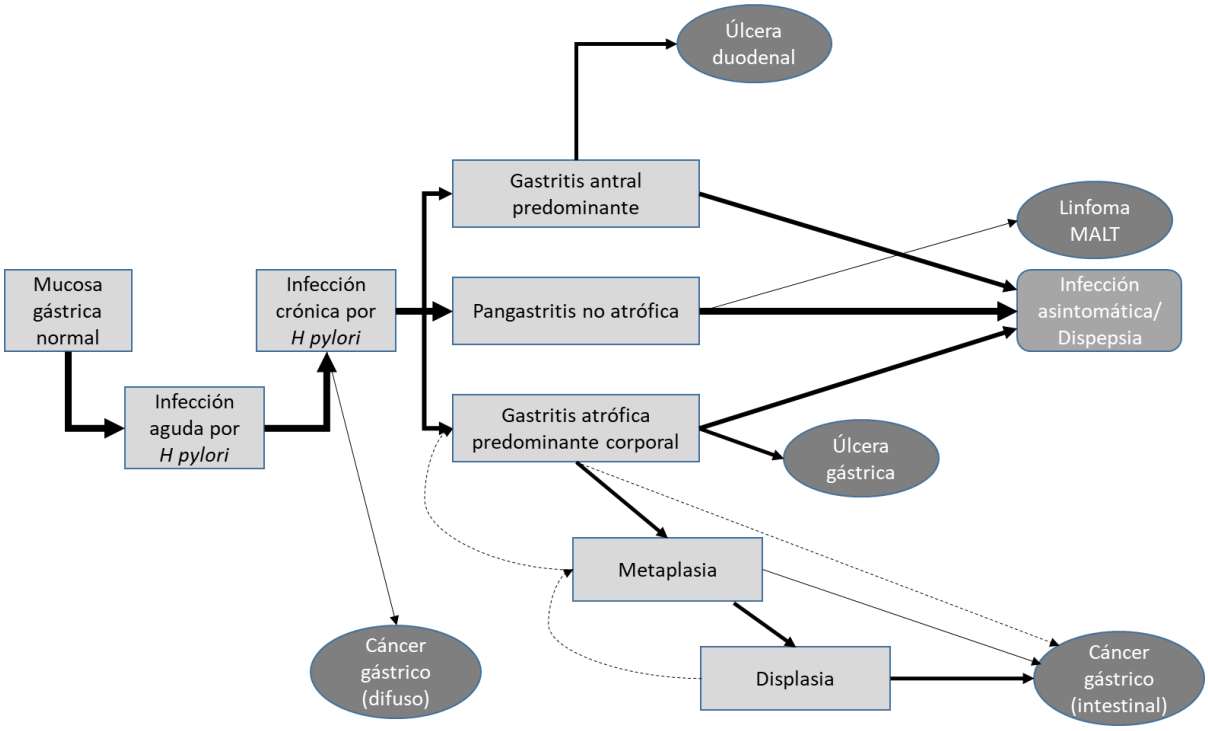


Figura 1: Historia natural de la infección por H pylori. Adaptado de Kandulski A et al (28)

Las vías más factibles de transmisión del H pylori son la fecal-oral y la oral-oral (29). Una vez la infección se establece tiende a permanecer durante toda la vida. Después de ingresar al estómago la bacteria inicia su proceso de colonización, el cual depende de factores como la motilidad mediada por su flagelo (que le da capacidad de ingresar a la capa mucosa), la producción de ureasa (que le posibilita sobrevivir en el ambiente con pH bajo del estómago), y la producción de adhesinas que le permite adherirse a las células epiteliales de la mucosa gástrica (ej. BabA, SabA, HopQ)(10).

El proceso carcinogénico inducido por el H pylori depende de la producción de toxinas por la bacteria y de la respuesta inflamatoria en el huésped. Durante y tras el proceso de colonización de la mucosa gástrica, el H pylori inicia la producción de toxinas incluyendo la proteína CagA (cytotoxin-associated gene A) y la toxina VacA (vacuolating cytotoxine A). CagA es traslocada a la célula huésped y posteriormente fosforilada por quinasas intracelulares, adquiriendo la capacidad de interactuar con múltiples vías de señalización celular, lo que desencadena alteraciones en el citoesqueleto y la función de la célula huésped (30). Las cepas bacterianas que producen CagA se han asociado a gastritis aguda, úlcera gástrica y cáncer gástrico. VacA tiene múltiples efectos incluyendo la creación de poros en la membrana celular que permiten la acción de la ureasa, la generación de vacuolas intracelulares, la muerte celular por apoptosis o necrosis, la

autofagia y la inhibición de los linfocitos T y B; sin embargo, la asociación de la expresión de VacA con cáncer gástrico continúa siendo controversial (10, 31).

La respuesta inflamatoria en el huésped también depende de la producción de toxinas por la bacteria (CagA la más relevante), pero la repuesta inmune tanto innata como adaptativa juegan un papel crucial en la activación de linfocitos y la secreción de citoquinas y quimiocinas proinflamatorias. Los componentes de la bacteria traslocados al interior de la célula desencadenan la activación de múltiples receptores, muchos de los cuales desembocan en la producción de IL8 y la consecuente atracción de neutrófilos y linfocitos, los que a su vez liberan proteínas quimiotácticas en el estómago que finalmente inducen inflamación aguda (10, 32). El *H pylori* genera también la respuesta adaptativa humoral y celular; sin embargo, esta respuesta generalmente no logra erradicar la infección, pues la escasa información disponible sugiere que solo un 15% de los sujetos infectados eliminan la infección de manera espontánea. El resultado de esto es una inflamación crónica de la mucosa gástrica que conduce a atrofia, hipo/acloridia y disbiosis, lo que favorece la transformación de las células hacia metaplasia intestinal, displasia y cáncer (10).

La razón por la cual la mayoría de personas permanecen asintomáticas durante el proceso de carcinogénesis no está claramente establecido. Algunos estudios de cohorte muestran que en promedio el 29% y máximo el 40% de los pacientes con diagnóstico de cáncer gástrico en estado temprano manifiesta síntomas asociados, particularmente durante el año previo al diagnóstico (33, 34); sin embargo, la modelación de datos a partir de la supervivencia observada luego del diagnóstico, sugiere que el periodo de latencia para el cáncer gástrico podría estar alrededor de 22 años (35).

De otra parte, un estudio que revisó el seguimiento endoscópico para cánceres de intervalo en pacientes mayores de 60 años mostró que con una frecuencia de seguimiento de dos años o menos, más del 90% de los casos podrían ser detectados en estados tempranos (36). En este estudio la mediana del tiempo de observación para cánceres sin hallazgo de lesiones precursoras en el seguimiento endoscópico fue 22 meses, mientras que para los casos con antecedente de lesiones precursoras fue 35 meses. Adicionalmente, el seguimiento de una cohorte de pacientes con diagnóstico de cáncer gástrico en estado temprano que no recibió tratamiento por diferentes motivos, mostró que la mediana de tiempo a la progresión hacia estados más avanzados fue 44 meses, con un 52% de casos diagnosticados por síntomas (34).

5.2 Detección temprana del cáncer gástrico

La detección temprana de cáncer gástrico se propone con abordaje diferencial dependiendo del nivel de riesgo poblacional (37, 38), incluyendo la tamización de base poblacional en poblaciones de alto riesgo (incidencia >20 por 100.000), la tamización selectiva para individuos de alto riesgo en poblaciones de riesgo medio (incidencia 10 a 20 por 100.000) y la vigilancia de lesiones precursoras (atrofia, metaplasia, displasia) en regiones de bajo riesgo (incidencia <10 por 100.000)(38). A pesar de la coherencia teórica de esta propuesta, en el momento solo Japón y Corea del Sur tienen programas nacionales de tamización y China realiza tamización solo en algunas regiones. Las pruebas utilizadas para la tamización son endoscopia y radiografía de doble contraste; sin embargo, la evidencia que soporta tales intervenciones es controversial. Un

metanálisis reciente basado en estudios observacionales mostró en el análisis por protocolo reducción de la mortalidad con la endoscopia (RR 0,52 IC95% 0,32-0,79), pero no con la radiografía de doble contraste (RR 0,80 IC95% 0,60-1,06), y no mostró eficacia con ninguna de las dos pruebas en el análisis por intención de tamizar (39).

La tamización selectiva se basa en la identificación de individuos de alto riesgo principalmente a partir de pruebas serológicas. Las pruebas más ampliamente usadas para este fin son el pepsinógeno sérico y la detección de H pylori, y con menor frecuencia la determinación de Gastrina-17. La información disponible sugiere que el uso combinado de estas pruebas tiene adecuado desempeño para la detección de atrofia gástrica (sensibilidad 74,7% y especificidad 95,6%)(40, 41); sin embargo, su uso como prueba de tamización requiere adecuado desempeño para la detección de otras lesiones precursoras y de cáncer invasor en estado temprano, en donde sus resultados no han sido favorables (42).

En la actualidad hay un gran auge investigativo para la búsqueda de nuevos biomarcadores en la detección temprana de cáncer gástrico incluyendo genes supresores de tumores, oncogenes, patrones de metilación, microARN, ARN largo no codificante, RNA circulante, células tumorales circulantes y ADN circulante (43). Una revisión reciente encontró 431 biomarcadores reportados en 149 estudios, pero solo 35 se han usado en fases más avanzadas de investigación, incluyendo el pepsinógeno como ya se anotó (44). La elevada heterogeneidad en el tipo de desenlaces utilizados, los diseños de los estudios, y la combinación de biomarcadores denotan el estado incipiente de la investigación en este campo.

El seguimiento de lesiones preneoplásicas se propone a partir de consensos de expertos y se basa en el riesgo de desarrollar cáncer gástrico en pacientes con esta condición (estudios observacionales). Las guías de Norteamérica y Europa recomiendan un abordaje estratificado en el que el riesgo se clasifica según la extensión y severidad de la lesión, y los antecedentes personales, a partir de lo cual recomiendan intervalos de seguimiento entre 3 y 5 años (37). Existe claro consenso acerca de las técnicas endoscópicas, el sistema de muestreo de la mucosa gástrica (Sidney) y el sistema de clasificación histopatológica de las lesiones preneoplásicas (OLGA/OLGIM); sin embargo, los intervalos de seguimiento no tienen un soporte de evidencia claro. Algunos estudios de historia natural sugieren que un seguimiento de dos años se asocia con mayor probabilidad de detectar la enfermedad en estados tempranos (36). Un consenso nacional, recomienda no hacer seguimiento en lesiones de bajo grado (OLGA/OLGIM 0, I, II), realizar seguimiento cada dos años en pacientes con lesiones extensas (antro y cuerpo con OLGA/OLGIM III, IV), y realizar seguimiento entre 6 meses y un año para pacientes con displasia de alto y bajo grado, respectivamente (45).

5.3 Modelos predictivos en detección temprana del cáncer

Aparte de las estrategias de tamizaje basadas en pruebas diagnósticas previamente descritas, la investigación sobre detección temprana de cáncer abarca otras estrategias como 1) la construcción de escalas de riesgo a partir de factores de riesgo establecidos, y 2) el cálculo de valores predictivos para síntomas individuales o agrupados.

5.3.1 Escalas de riesgo a partir de factores de riesgo establecidos

Las escalas de riesgo pretenden hacer una mejor selección de los pacientes a tamizar. Una revisión sistemática de 2022 identificó 12 estudios que desarrollaron modelos predictivos para cáncer gástrico (46). De los estudios identificados solo uno se realizó en Estados Unidos, los restantes corresponden a poblaciones asiáticas; solo uno utilizó técnicas de aprendizaje de máquinas y todos se basan en estudios epidemiológicos (corte transversal, casos y controles y cohortes). Los tamaños muestrales oscilan entre 380 y 1.370.000 sujetos. Entre todos los modelos se identificaron 30 predictores diferentes que incluyen variables sociodemográficas, variables clínicas, de estilo de vida y de laboratorio. En total 13 de los predictores se identificaron en dos o más modelos que corresponden a 1) variables sociodemográficas: edad, sexo, 2) estilos de vida: consumo elevado de sal, tabaquismo, alcohol, sedentarismo, e índice de masa corporal, regularidad de las comidas, 3) variables de laboratorio: *H pylori*, pepsinógeno, hemoglobina, gastrina-17, y 4) antecedentes familiares de cáncer gástrico.

Una segunda revisión sistemática de 2023 identificó 71 estudios que desarrollaron modelos predictivos en cáncer gástrico (17). Solo 8 de los 71 estudios se realizaron en poblaciones no orientales (Estados Unidos 2, Europa 5, Irán 1), 33 fueron desarrollados a partir de características moleculares de datos recolectados en biobancos y uno se basó en poblaciones de ensayos clínicos aleatorizados sobre tratamiento de cáncer gástrico. Los factores predictivos identificados incluyen variables sociodemográficas, variables de laboratorio (incluyendo marcadores tumorales), estilos de vida, historia familiar y datos “ómicos” (genómica, transcriptómica, proteómica, metabolómica). Estos últimos, al igual que algunas variables de laboratorio están orientados al pronóstico de la enfermedad y no a su predicción; de hecho, los modelos de predicción solo representan el 45% de los estudios incluidos en la revisión y en su mayoría son aplicables a sujetos que ya están en vigilancia endoscópica.

En una búsqueda adicional, se identificaron 4 publicaciones recientes no incluidas en las revisiones sistemáticas descritas: dos de Irán, una de Japón y una de China (Hong Kong)(18, 20, 47, 48). Los dos estudios iraníes se basan en cohortes de pacientes dedicadas al estudio de cáncer gástrico, y a su vez, el estudio japonés y el estudio chino se basan en registros clínicos, el primero de una base de datos endoscópica y el segundo de registros hospitalarios electrónicos. En total los modelos usaron entre 19 y 28 variables con grandes diferencias entre ellos. Affrash MR et al (Irán)(18), orientaron el análisis hacia factores relacionados con el estilo de vida; sin embargo, incluyeron variables de laboratorio, antecedentes familiares y algunos síntomas asociados al cáncer gástrico. Leung WK et al (China)(48), orientaron su modelo a pacientes que habían sido tratados para erradicación de *H pylori* e incluyen en su modelo 15 condiciones de comorbilidad (4 diagnósticos relacionados con enfermedad gástrica) y 6 medicamentos. Taninaga J et al (Japón)(47), incluyen en su modelo 10 parámetros de laboratorio y 4 diagnósticos endoscópicos. La edad, el sexo, y el índice de masa corporal (IMC) son considerados en los 4 modelos; el tabaquismo, el consumo de alcohol y el estado o antecedente de *H pylori* en 3 de ellos; y la comorbilidad común a dos de los 4 modelos (excepto Japón) incluye cardiopatía, hipertensión, insuficiencia renal, accidente cerebrovascular y enfermedad hepática. En total 11 diagnósticos de comorbilidad fueron considerados por Mohammadnezhad K et al y Leung WK et al.

Adicionalmente, dos estudios comparan las escalas de riesgo construidas mediante métodos epidemiológicos convencionales (modelos multivariantes), con modelos basados en aprendizaje de máquinas. Uno de los estudios utiliza la edad y el sexo, síntomas asociados al cáncer gástrico (15 variables) y pruebas de laboratorio (7 variables)(19). El otro estudio utiliza 17 variables en total incluyendo edad, sexo, raza, condición de inmigrante, IMC, infección por *H pylori,* antecedentes familiares, antecedentes personales relacionados con cáncer gástrico y 5 condiciones de comorbilidad (21). Briggs E et al encontraron mejor desempeño de los modelos de aprendizaje de máquinas medido por el área bajo la curva (0,81 para el modelo convencional vs 0,86-0,87 para los modelos LASSO, máquina de vectores de soporte, bosques aleatorios y potenciación de gradiente extremo); sin embargo, la medida de precisión global (*accuracy*) no mostró mayores diferencias (0,87 a 0,89). A su vez, Huang RJ et al no encuentran diferencia en la precisión (*accuracy*) entre la predicción basada en un modelo de regresión logística convencional y lo modelos de aprendizaje de máquinas (0,72 IC95% 0,69-0,76 para el modelo convencional vs 0,72-0,78 para los modelos LASSO, máquina de vectores de soporte, bosques aleatorios y vecinos mas cercanos o KNN por su sigla en inglés).

5.3.2 Escalas de riesgo a partir de síntomas

Diagnosticar la enfermedad a partir de síntomas se dificulta por factores de tipo sociocultural como el estigma y el nivel educativo, y por barreras de los servicios de salud como la capacidad del personal de cuidado primario para adelantar un proceso diagnóstico adecuado (49).

No siempre resulta fácil relacionar los síntomas con la enfermedad. Esfuerzos para mejorar la capacidad diagnóstica a partir de síntomas incluyen campañas educativas, guías clínicas, diagnóstico oncológico centralizado, etc., sin que a la fecha se haya identificado una intervención que de forma sostenida mejore la capacidad diagnóstica del cáncer a partir de la expresión de síntomas más allá de la mayor capacidad de respuesta y mejor organización del sistema de salud como determinante general del proceso salud-enfermedad.

El cálculo de valores predictivos para síntomas específicos se ha realizado para diferentes tipos de cáncer (14). Si bien en la mayoría de casos el valor predictivo ha sido bajo, en algunas circunstancias se han encontrado valores predictivos positivos similares a los reportados para algunas pruebas de tamizaje de cáncer. Algunas limitaciones de tales estudios incluyen un número reducido de sujetos, un tiempo de observación limitado, y su concentración en la correlación de síntomas sin considerar algoritmos de decisión diagnóstica apropiados.

La disponibilidad de grandes bases de datos ha permitido el desarrollo de algoritmos que superan varias de las limitaciones descritas. En un trabajo adelantado por la University Park en Notingham-Inglaterra, se desarrollaron algoritmos para diferentes tipos de cáncer incluyendo los de estómago, colon, pulmón, y el cáncer gastroesofágico (Tabla 1)(50). Los algoritmos fueron desarrollados a partir de la información de las bases de datos del sistema nacional de salud en el Reino Unido y utilizaron un análisis de cohortes para la identificación de síntomas asociados a la enfermedad y la construcción de los modelos diagnósticos. Como se observa en la tabla, los algoritmos obtienen una

adecuada sensibilidad, pero un valor predictivo positivo bajo a pesar de tratarse de síntomas compatibles con estado avanzado de la enfermedad, lo que incluye en el caso de cáncer gástrico disfagia, hematemesis y pérdida de peso.

Tabla 1. Características de algoritmos diagnósticos desarrollados para la detección de cáncer en cuidado primario (50)

Tipo de cáncer	Síntomas	Factores de riesgo	Algoritmo (PC10%)	
			Sensibilidad	VPP
Cáncer gastroesofágico	Disfagia, hematemesis, dolor abdominal, pérdida de apetito, pérdida de peso, anemia	Tabaco, alcohol, antecedente familiar	76,5	1,2
Cáncer de pulmón	Hemoptisis, tos en el último año, pérdida de apetito, pérdida de peso, anemia	Tabaquismo, diagnóstico previo de cáncer (diferente de pulmón), EPOC, índice de pobreza	77,3	1,3
Cáncer de páncreas	Disfagia, distensión abdominal, constipación en el último año, dolor abdominal, pérdida de apetito, pérdida de peso	Tabaquismo, antecedente de diabetes tipo 2, antecedente de pancreatitis crónica	62,4	0,6
Cáncer colorrectal	Rectorragia, cambio en el hábito intestinal en el último año, dolor abdominal, pérdida de apetito, pérdida de peso, anemia	Alcohol, antecedente familiar de cáncer gastrointestinal	70,6	1,5

En otros escenarios se ha evaluado la capacidad de respuesta de los servicios de cuidado primario a partir de la presentación de síntomas no específicos (“non-alarm symptoms”), organizando servicios multidisciplinarios de diagnóstico (51). Sin embargo, la información disponible al momento sugiere un bajo desempeño y un elevado costo y complejidad de estos servicios, por lo que se requieren estudios para la construcción de algoritmos diagnósticos con mejor desempeño predictivo.

A diferencia de las escalas construidas a partir de factores de riesgo, las escalas basadas en síntomas tendrían menor influencia del riesgo poblacional. A pesar de ello, los espectros epidemiológicos varían entre poblaciones lo que hace que la capacidad discriminativa de un síntoma se modifique en razón a la prevalencia de enfermedades que constituyen diagnósticos diferenciales del cáncer, aspecto no tenido en cuenta en la construcción de este tipo de escalas.

Algunos estudios proponen la búsqueda de cáncer en pacientes que presentan condiciones específicas asociadas al cáncer como la hipercoagulabilidad y la inmunodeficiencia. Un metanálisis reciente encontró que un tamizaje básico en pacientes que presentan trombosis sin factor desencadenante aparente detecta cáncer oculto en uno de cada veinte individuos, siendo la historia clínica y el examen físico el principal método de detección (52). El abordaje representa un enfoque diferente por dirigirse a la búsqueda de cáncer en general a partir de una expresión sintomática común a todos los tipos de cáncer. El análisis presentado establece que el 70% de los casos se detectó con

base en la historia clínica y el examen físico y no se encontró diferencia con el uso de pruebas especializadas como la PET-CT frente al uso de pruebas rutinarias (radiografía de tórax, hemograma, etc.). De menara similar a la trombosis asociada a cáncer, la expresión clínica de la inmunosupresión no es muy frecuente en enfermedad temprana, lo que hace que las tasas de detección a partir de ella, no sea sustancialmente altas en la población general de pacientes con cáncer.

Un reto para este abordaje lo representa la manifestación de síntomas del cáncer, la cual podría requerir un esfuerzo diferencial dependiendo de la cantidad de diagnósticos diferenciales asociados. En tal sentido Koo MM y colaboradores proponen una taxonomía para los diferentes tipos de cáncer, útil a la construcción de algoritmos diagnósticos desde esta perspectiva (Cuadro 1)(53).

Cuadro 1. Tipos de cáncer según expresión sintomática. Tomado de Koo MM et al (53)

Tipos de cáncer con limitada expresión sintomática de alto valor predictivo
Mama
Vejiga
Melanoma
Tiroides
Testículo
Tipos de cáncer con amplia expresión sintomática de valor predictivo variable
Colon y recto
Estómago
Pulmón
Orofaringe
Ovario
Riñón
Tipos de cáncer con amplia expresión sintomática de bajo valor predictivo
Sistema nervioso central
Neoplasias hematológicas

5.4 Consideraciones metodológicas

5.4.1 Uso de bases de datos administrativas

De forma ideal, la identificación y cuantificación del riesgo de cáncer (desenlace) asociado a una manifestación clínica (variable predictiva) debería hacerse en un estudio prospectivo; sin embargo, un estudio convencional con el tamaño de muestra suficiente para detectar una enfermedad de baja frecuencia a partir de síntomas frecuentes en la población general, resulta poco factible de realizar en el escenario del cuidado primario. Adicionalmente, este diseño requiere de un listado de síntomas a priori, lo que implica

que cualquier motivo de consulta no incluido en el listado sería obviado del análisis independientemente de su potencial predictivo (14).

La existencia de grandes bases de datos ha abierto la posibilidad de realizar análisis que tienen base poblacional, permiten el seguimiento prospectivo, eliminan el sesgo de la selección a priori de las variables a incluir, mejoran la posibilidad de estudiar la asociación entre exposiciones y enfermedades de baja prevalencia, y permiten el análisis de múltiples dominios en una misma cohorte de sujetos y el análisis de patrones de utilización de servicios en enfermedades crónicas (54). Sin embargo, el uso de dichas bases de datos tiene limitaciones debido a que corresponden usualmente a datos recopilados con fines administrativos y no con fines epidemiológicos, por lo que la calidad de los mismos es con frecuencia deficiente.

La calidad de la información es un concepto complejo que incluye múltiples dominios dependiendo del marco analítico que se utilice y el propósito de utilización de los mismos (55). En general, el aporte limitado de información existente en las bases de datos administrativos en aspectos como los determinantes de la salud y los factores de riesgo (prácticamente inexistente), se compensa en gran parte por fortalezas como el seguimiento integral de los individuos, la recolección continua de los datos y la medición objetiva de las variables centrales para la administración (54).

Las limitaciones descritas han inducido el desarrollo de métodos de análisis robustos incluyendo mecanismos de validación de los resultados obtenidos, procesos que deben observarse con detenimiento en la construcción de propuestas basadas en estas fuentes de información (56). En resumen, a pesar de las desventajas mencionadas, el uso de bases de datos administrativas interconectadas ofrece un recurso poderoso para la investigación en salud pública y en consecuencia el auge de la investigación con este tipo de fuentes ha promovido un desarrollo metodológico con mejora progresiva en la calidad de los estudios tanto del área de la salud como de las áreas sociales en general.

Los tipos de datos usualmente disponibles en estas fuentes incluyen la utilización de servicios de salud, registros de población, ubicación de la residencia, vínculos familiares, resultados educativos y uso de servicios sociales. En el caso de Colombia, las bases de datos disponibles recogen información relacionada con aspectos como el proceso de facturación de medicamentos y procedimientos, datos de reclamación de las aseguradoras, desenlaces en salud, eventos adversos, terapias instauradas y consumo de medicamentos.

5.4.2 Modelos de predicción clínica

Los modelos de predicción utilizan un conjunto de variables para estimar la probabilidad absoluta de que un evento esté presente o vaya a ocurrir en un lapso de tiempo dado (57). Las variables predictoras utilizadas en la construcción de modelos comprenden una amplia que gama que incluye características sociodemográficas, factores de riesgo, historial de enfermedad, estudios clínicos (laboratorio, imágenes diagnósticas, histopatología, etc.), y marcadores moleculares o genéticos.

En tal sentido, los modelos de predicción pueden utilizarse para diferentes propósitos como orientar la modificación de comportamientos de riesgo, guiar decisiones terapéuticas y estratificar los pacientes para hacer valoración o seguimiento ajustado a su perfil de riesgo (58). Además de tener un desempeño clínico adecuado, los modelos

de predicción deberían ser simples y fáciles de usar, ya que, si el modelo es difícil de implementar su probabilidad de utilización será baja. Algunos aspectos que dificultan la implementación de los modelos incluyen la necesidad de información difícil de obtener, un elevado consumo de tiempo, altos costos o el uso en individuos o población inapropiada (falta de validación)(59).

En términos generales, el proceso de construcción de un modelo implica una adecuada definición de su objetivo, la obtención de datos para la modelación, la construcción del modelo, y su validación (60). El objetivo determina no solo el tipo de abordaje (por ejemplo, relación de corte transversal si el propósito es diagnóstico o relación longitudinal si el propósito es determinación de riesgo), sino además la población en la que se espera hacer la predicción y por ende el tipo de datos requeridos para la modelación. Por ejemplo, la estimación de la probabilidad de tener cáncer ante la presencia de un factor dado puede ser diferente si el modelo predictivo se construye a partir de datos de pacientes asistiendo a una clínica especializada entre quienes hay una alta prevalencia de la enfermedad, o si se basa en información de población que asiste a centros de cuidado primario en donde la prevalencia de la enfermedad es menor.

La disponibilidad de datos apropiados es uno de los mayores retos en la construcción de modelos predictivos. De forma ideal, los datos deberían recolectarse de forma prospectiva y con una definición apropiada de las variables a incluir en el modelo; sin embargo, tal información raramente se puede obtener por lo que lo más frecuente es el uso de datos pre-existentes y recolectados para un propósito diferente de la construcción del modelo, incluyendo datos de ensayos clínicos, estudios de cohorte, registros de enfermedad y bases de datos administrativas. En consecuencia, previo a la construcción del modelo debe haber un proceso de evaluación de la calidad de la información y una preparación de los datos, que puede incluir, entre otros, etiquetado, recategorización de variables y manejo de valores ausentes mediante técnicas de imputación (57).

La construcción de los modelos de predicción clínica parte de la identificación de potenciales predictores del desenlace objeto de estudio, lo cual tradicionalmente se ha basado en la consideración teórica de la asociación entre dicho predictor y el desenlace correspondiente y en la disponibilidad de datos para los predictores identificados de forma preliminar. La selección final de predictores a incluir en el modelo sigue parámetros estadísticos bien definidos, al igual que los procesos de validación interna y externa del modelo (57, 60, 61).

Como se describió previamente, existen múltiples modelos predictivos en el área de cáncer desde el establecimiento de riesgo de la enfermedad hasta el cuidado al final de la vida (62). Mas allá de la identificación de condiciones de riesgo, la predicción es un objetivo permanente en el cuidado oncológico en donde la ausencia de síntomas durante el periodo de latencia de una enfermedad con elevada letalidad confiere gran relevancia al ejercicio predictivo para la detección temprana de la misma; pero adicionalmente, la predicción de su evolución dependiendo de las condiciones al momento del diagnóstico determina el plan de tratamiento y el esfuerzo terapéutico para su manejo.

En respuesta a la incertidumbre y la necesidad de predicción en torno al manejo del cáncer, existe hoy una gran cantidad de herramientas que tendrían en teoría, el potencial de mejorar los resultados en el manejo de pacientes oncológicos a partir del

mejoramiento en la consistencia y la calidad de las decisiones clínicas, lo que a su vez podría resultar en una mejor utilización de recursos para el control de la enfermedad. La evidencia disponible sugiere gran potencial de aplicación clínica y en efecto muchas de estas herramientas se utilizan en la rutina de los servicios para la toma de decisiones, incluyendo una tendencia creciente de las herramientas que incorporan datos moleculares como se anotó previamente. Sin embargo, existe gran incertidumbre sobre los mecanismos apropiados para su implementación y el impacto que dichas herramientas pueden tener una vez introducidas en la rutina de los servicios (63); por ello, el ejercicio de validación es solo el inicio en el proceso de desarrollo, ya que este debería complementarse posteriormente con la evaluación de su aceptabilidad, el impacto económico de su implementación y la valoración de potenciales efectos adversos como la sobre-medicalización derivada de los modelos de riesgo.

5.4.3 Aprendizaje de máquinas

El aprendizaje de máquinas es una forma de inteligencia artificial y se refiere a la programación de computadores de forma que generen aprendizaje automatizado acerca de un problema a partir de un conjunto de datos, lo que permite luego hacer algún tipo de inferencia (64). Existen diferentes clasificaciones de los sistemas de aprendizaje de máquinas, pero en términos generales, el aprendizaje de máquinas se hace por métodos supervisados o no supervisados. En el primer caso se utiliza un conjunto de datos de prueba etiquetado para definir las variables independientes utilizadas a fin de predecir el desenlace esperado. En el aprendizaje no supervisado los datos no tienen etiquetas y no se conoce el desenlace durante el proceso de aprendizaje, por lo que el modelo de aprendizaje definirá las variables independientes a partir de descubrir relaciones entre atributos (64). También existen sistemas semi-supervisados que combinan datos etiquetados y no etiquetados, siendo usualmente los últimos los de mayor cantidad en este tipo de sistemas. Los modelos de aprendizaje de máquinas también se han categorizado dependiendo de su capacidad de aprendizaje incremental o fija (en línea y fuera de línea) o si se basan en características fijas o en modelos complejos.

El aprendizaje de máquinas resulta útil para la solución de problemas cuyo abordaje requiere gran cantidad de ajustes manuales o un gran esfuerzo de parametrización para la búsqueda de soluciones a problemas que no ha sido posible resolver por las técnicas tradicionales, para soluciones que deben aplicarse en ambientes fluctuantes, y para obtener información sobre problemas complejos y grandes cantidades de datos (65).

Los principales retos del aprendizaje de máquinas son la idoneidad de los datos y la selección apropiada de algoritmos de aprendizaje. La idoneidad de los datos puede verse afectada por insuficiencia de la información suministrada para el aprendizaje (falta de variables para la predicción), por falta de representatividad de los datos para realizar las inferencias correspondientes, por mala calidad (errores en el registro o en el etiquetado, valores faltantes, etc.) o por exceso de información irrelevante que genera distorsiones en el proceso de aprendizaje (65).

Los principales problemas en la selección del modelo son el sobreajuste y la falta de ajuste. El sobreajuste puede ocurrir cuando el proceso de aprendizaje genera un modelo demasiado complejo que se ajusta muy bien a los datos de entrenamiento, pero no se reproduce en los datos de validación o aplicación. El sobreajuste podría corregirse

simplificando el modelo mediante la reducción de parámetros incluidos, regularizando el modelo mediante la reducción de grados de libertad e interacción entre variables, ampliando el conjunto de datos de entrenamiento (más observaciones) o mejorando la calidad de los datos (retirando valores atípicos, complementando valores ausentes, etc.). La falta de ajuste es lo opuesto al sobreajuste y puede ocurrir con modelos demasiado simples (falta de parámetros), por sobre regularización o por la inclusión de parámetros inapropiados (65).

Existen diversos métodos de aprendizaje de máquinas que se han utilizado en el desarrollo de modelos predictivos en diversas áreas de la salud. En una revisión sistemática que identificó 71 estudios de predicción o pronóstico de cáncer gástrico, los métodos más utilizados fueron: para la selección de parámetros LASSO (Least Absolute Shrinkage and Selection Operator), y para la discriminación bosques aleatorios, máquinas de vectores de soporte y redes neuronales (aprendizaje profundo); sin embargo, en publicaciones recientes se utiliza con creciente frecuencia el aumento o potenciación del gradiente (Gradient Boosting)(17).

LASSO es un método de regularización que combina un modelo de regresión con un procedimiento de contracción de los coeficientes. Debido a esta restricción, LASSO hace que los coeficientes de menor significancia lleguen a valor cero y, por tanto, proporciona modelos interpretables con mayor facilidad para la selección de subconjuntos de variables, mayor precisión en la capacidad predictiva, y mayor estabilidad (pequeños cambios en la variable no producen grandes cambios en la predicción)(17, 66, 67).

Un bosque aleatorio es un conjunto de clasificadores compuestos por árboles de decisión que se generan utilizando dos fuentes de aleatorización. Primero, se entrena cada árbol de decisión individualmente en una muestra aleatoria con reemplazo de los datos originales con el mismo tamaño que el conjunto de entrenamiento. Una segunda fuente de aleatorización en los bosques aleatorios es el muestreo de atributos, para eso, en cada nodo de división se selecciona aleatoriamente un subconjunto de las variables de entrada para buscar la mejor división. En general se crean múltiples clasificadores débiles, pero la predicción final se da por mayoría de votos generando un resultado de alta precisión (17, 68).

A su vez, la potenciación del gradiente tiene como fundamento resolver problemas de aprendizaje de máquinas combinando clasificadores débiles escogidos de forma cuidadosa para formar un conjunto cuyas predicciones sean más acertadas. La potenciación parte de un algoritmo de aprendizaje débil con la meta de mejorar su desempeño. Para ello se aplica sucesivamente el algoritmo débil a versiones modificadas del conjunto de entrenamiento, de este modo en cada iteración se genera un nuevo clasificador débil que posteriormente se combina formando un clasificador mejorado (68, 69).

Las máquinas de vectores de soporte se basan en modelos matemáticos que pretenden identificar el máximo margen posible entre las fronteras de decisión que separan dos categorías (65). De forma similar a las redes neuronales, el método se utiliza en la predicción de la ocurrencia de enfermedad y el pronóstico de la misma (64).

Las redes neuronales artificiales activan un desenlace frente a diferentes parámetros o combinaciones de las variables independientes, lo cual pueden hacer en una o en varias

capas de relaciones que le confieren capacidad de aprendizaje; en esa medida han mostrado ser un buen método de clasificación (70). Este método de aprendizaje de máquinas ha sido utilizado en la predicción de la ocurrencia, progresión y recurrencia del cáncer de diversas formas. En la predicción de la ocurrencia de cáncer se ha usado para la interpretación de imágenes diagnósticas, el análisis de información histopatológica y molecular, el análisis de información sobre factores de riesgo y el análisis combinado de los tipos de datos mencionados (70-72).

6. Hipótesis

Hipótesis alterna: las trayectorias de atención registradas en el sistema de salud permiten identificar pacientes que serán diagnosticados con cáncer gástrico en un periodo de tiempo de uno a dos años.

Hipótesis nula: las trayectorias de atención registradas en el sistema de salud NO permiten identificar pacientes que serán diagnosticados con cáncer gástrico en un periodo de tiempo de uno a dos años.

7. Diseño metodológico

7.1 Diseño general del estudio

El estudio incluyó dos componentes principales:

1. Construcción de un modelo multivariado de predicción clínica
2. Construcción de modelos de predicción basados en aprendizaje de máquinas

El trabajo se desarrolló partir de las bases de datos disponibles en el sistema de salud y tuvo como etapas principales

- Limpieza de la base de datos
- Validación del etiquetado
- Desarrollo de un modelo multivariante
- Desarrollo de modelos de aprendizaje de máquinas

El estudio fue aprobado por el Comité de Ética de la Facultad de Medicina de la Universidad Nacional de Colombia y por el Comité de Ética e Investigaciones del Hospital Universitario San Ignacio – Facultad de Medicina de la Pontificia Universidad Javeriana. en Bogotá.

Para el análisis se utilizaron los paquetes estadísticos Stata® 16.0 (estadística descriptiva, medidas de asociación), R© 4.3.1 (modelos multivariantes, riesgos competitivos), Python™ 3.12.0 (integración de bases de datos, aprendizaje de máquinas), y el marco de trabajo Apache Spark™ 3.4.1.

De Spark se utilizó la librería para aprendizaje de máquinas (MLlib) y el módulo para bases de datos (SQL Spark). La integración de Spark y Python se hizo mediante PySpark. En Python se usaron las librerías Pandas (administración e integración de bases de datos), Numpy (generación de matrices con estructura universal), y Random (generación de números aleatorios en la organización de base de datos de entrenamiento y validación). Los datos fueron trabajados en un servidor institucional con 48 procesadores y en Google Colab.

Para la elaboración del informe se utilizó el paquete de Microsoft Office® 2019 en las aplicaciones Excel y Power Point para la preparación de gráficas y figuras, y Microsoft Word como procesador de texto.

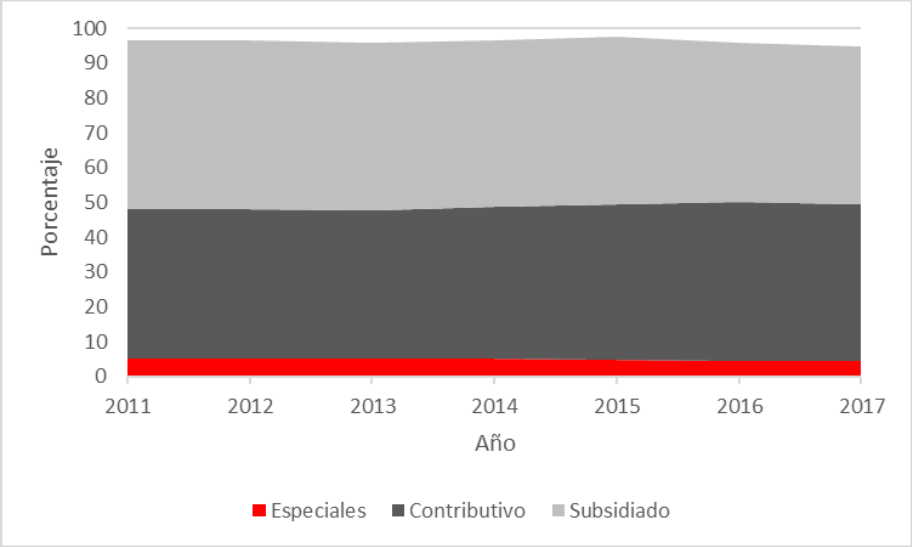
7.2. Fuentes de información

Se utilizó la base de datos para cálculo de la Unidad de Pago por Capitación en el régimen contributivo del sistema de salud de Colombia, la información de defunciones del país proveniente del sistema de estadísticas vitales y la base de datos del Registro Poblacional de Cáncer de Bucaramanga (73).

El sistema de salud colombiano se basa en el aseguramiento individual y tiene dos regímenes principales de aseguramiento: 1) el régimen contributivo al que deben afiliarse todos los asalariados o pensionados y los trabajadores independientes con ingreso superior a un salario mínimo y 2) el régimen subsidiado que se encarga del aseguramiento de las personas sin capacidad de pago y no cubiertas por el régimen contributivo. En el primer caso el seguro cubre el primer beneficiario y a sus dependientes. Los dos regímenes cubren los servicios sanitarios dentro de un paquete de beneficios definido por el Ministerio de Salud y Protección Social, pero administrado por compañías de seguros (74).

Para el año 2022 la cobertura de aseguramiento en el país correspondía al 99,1% del total de la población y el régimen contributivo cubría el 45,3% de la población colombiana (75). La evolución de la cobertura de afiliación durante el periodo de estudio se muestra en el Gráfico 1.

Gráfico 1. Tendencia de la cobertura de aseguramiento según régimen de seguros durante el periodo de estudio en Colombia



Los regímenes especiales corresponden a poblaciones con sistema de seguros independiente como las fuerzas armadas y los maestros. Fuente (75)

Para brindar los servicios de salud, toda compañía de seguros de salud recibe un pago anual per cápita por afiliado, y para definir el valor anual de la unidad de pago per cápita (UPC), las aseguradoras del régimen contributivo deben suministrar anualmente al Ministerio de Salud y Protección Social información detallada sobre la prestación de

servicios de salud para cada afiliado (76). Los datos son reportados de forma estructurada y constituyen una base de datos con millones de registros (base de datos para el estudio de la suficiencia UPC - BDUPC).

Cada atención en salud dentro de la BDUPC se caracteriza con las siguientes variables:

1. Asegurador (Entidad Promotora de Salud o Entidad Administradora de Plan de Beneficios según la denominación vigente)
2. Identificador (Corresponde a un código encriptado que reemplaza la cédula de ciudadanía a fin de anonimizar los datos, pero se mantiene constante en todos los periodos analizados y en la base de datos de defunciones)
3. Fecha de nacimiento
4. Sexo
5. Municipio en donde se presta la atención (Código DIVIPOLA)
6. Diagnóstico relacionado (Clasificación Internacional de Enfermedades - CIE10)
7. Fecha de atención o prestación del servicio
8. Tipo de atención: medicamento o procedimiento (diagnóstico o terapéutico)
9. Código de procedimiento (Clasificación Única de Procedimientos en Salud – CUPS)(77)
10. Código de medicamento (Código Único Nacional de Medicamentos -CUM- ó Clasificación anatómico, terapéutico, químico -ATC-)(78, 79)
11. Ámbito del procedimiento (Ambulatorio, hospitalario, urgencias, domiciliario)
12. Forma de reconocimiento o pago del procedimiento (Capitación, evento, grupo relacionado de diagnóstico, autorizado, directo)
13. Días de estancia
14. Valor pagado
15. Código de descuento (copago o cuota moderadora)
16. Identificador del prestador (Código de habilitación según el Registro especial de prestadores de servicios de salud – REPSS)

La información de defunciones se obtuvo a partir del registro único de afiliados en su módulo de estadísticas vitales (RUAFND: nacimientos y defunciones)(80), el cual contiene en relación con la mortalidad información codificada sobre:

1. Identificador (ID)
2. Sexo
3. Edad
4. Nivel educativo
5. Causas de muerte
6. Lugar de defunción
7. Ciudad y departamento de defunción

La base de datos del Registro Poblacional de Cáncer de Bucaramanga contiene mayor detalle de los casos de cáncer diagnosticados y residentes en la ciudad de Bucaramanga incluyendo identificador, fecha de diagnóstico, localización anatómica, sub-localización anatómica, tipo histológico, grado de diferenciación. Esta información se utilizó para el análisis de sensibilidad de los algoritmos informáticos para el etiquetado de casos.

7.3 Definición de sujetos de estudio

Se incluyeron en el análisis pacientes mayores de 30 años y menores de 90 años, afiliados al régimen contributivo de seguridad social en Colombia, quienes recibieron alguna atención por los servicios de salud del país desde el año 2011 hasta el año 2017, y registrados en la base de datos de los estudios de suficiencia de la Unidad de Pago por Capitación del Ministerio de Salud (BDUPC)(76). Para la construcción del modelo multivariante la edad mínima para inclusión en el análisis fue 40 años.

7.4 Limpieza de las bases de datos y selección de periodos de análisis

Debido a las deficiencias de las bases de datos administrativas descritas como parte de los antecedentes teóricos, se realizó un análisis de la información disponible a fin de determinar la suficiencia de los mismos (*completness*), su comparabilidad y su consistencia. A partir de este ejercicio se seleccionaron los periodos a analizar dentro de la información disponible (2011 a 2017), se completaron datos faltantes y se eliminaron registros con valores ausentes en los que no se puedo completar la información.

La disponibilidad de datos para cada año y cada variable se presenta en la Tabla 2. El sexo y la fecha de nacimiento tuvieron valores faltantes (en blanco) en todos los años y para los años 2011 y 2015 la variable fecha de nacimiento solo reporta el año de nacimiento (Marcado *). Los códigos de medicamentos para el año 2013 corresponden al Código Único Nacional de Medicamentos (CUMS)(78) y para 2016 a códigos del Sistema de Clasificación Anatómica, Terapéutica, Química (ATC por su sigla en inglés)(79).

Tabla 2. Disponibilidad de datos discriminada por año en la base de datos

Variable \ Año	2011	2012	2013	2014	2015	2016	2017
Aseguradora	X	X	X	X	X	X	X
Identificación	X	X	X	X	X	X	X
Fecha nacimiento	X*	X	X	X	X*	X	X
Sexo	X	X	X	X	X	X	X
Municipio	X	X	X	X	X	X	X
Diagnóstico	X	X	X	X	X	X	X
Fecha de atención	X	X	X	X	X	X	X
Tipo servicio	X	X	X	X	X	X	X
Procedimiento	X	X	X	X	X	X	X
Nombre del medicamento	X	-	X	X	X	-	X
Código del medicamento	-	-	X	-	-	X	-
Ámbito del procedimiento	X	X	X	X	X	X	X
Forma de reconocimiento	X	X	X	X	X	X	X
Días de estancia	X	X	X	X	X	X	X
Valor pagado	X	X	X	X	X	X	X
Tipo código descuento	X	X	X	X	X	X	X
Código habilitación del prestador	X	X	X	X	X	X	X
Razón social del prestador	-	-	-	X	-	X	X

Adicional a los datos faltantes (en blanco), las variables fecha de nacimiento y sexo reportan para el año 2012 un porcentaje mayor al 40% con valor no interpretable ("\N").

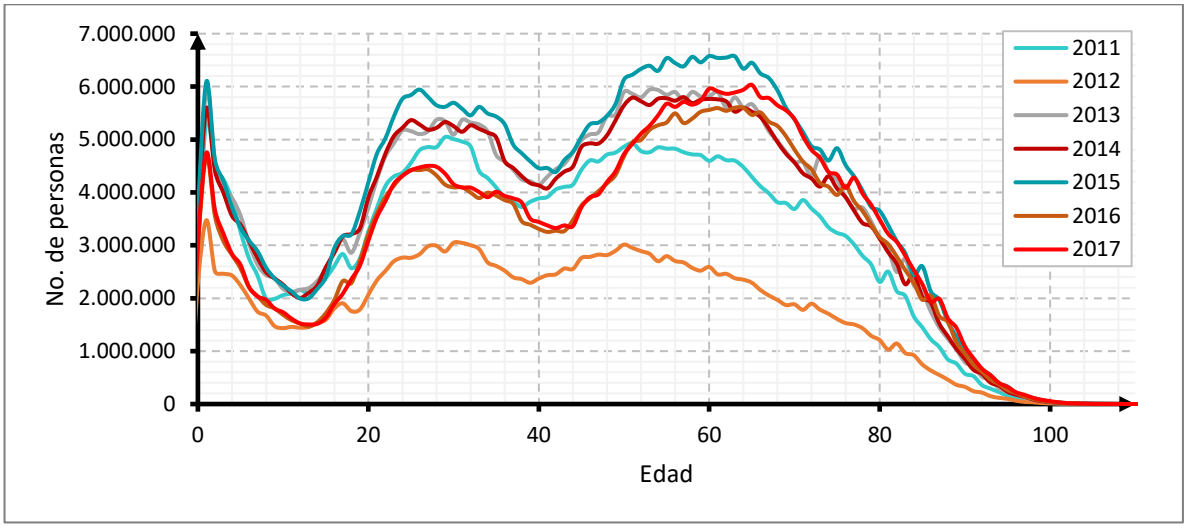
Tabla 3. Registros totales y registros faltantes para fecha de nacimiento y sexo discriminados por año en las bases de datos disponibles

Año	No. total de registros	Suma de IDs únicos por mes (12 meses)	Faltantes fecha de nacimiento	Faltantes sexo
2011	330.820.973	65.072.941	135.394	1.267.026
2012	344.705.008	66.434.458	12.367	15.299
2013	386.780.187	68.428.141	27.792	36.341
2014	383.174.082	65.436.973	77.669	1.279.749
2015	422.854.670	69.013.982	99.77	1.817.100
2016	338.071.694	61.578.816	113.33	1.616.129
2017	352.211.168	64.978.325	168.516	1.998.998

La suma de IDs únicos para 12 meses corresponde a la suma de IDs únicos mes a mes; es decir que un sujeto puede estar repetido en meses diferentes del año, pero no en el mismo mes.

La revisión del comportamiento de la variable edad mostró una distribución similar para la mayoría de años analizados, excepto para el año 2012 como se observa en el Gráfico 2. Como se anotó, el año 2012 fue el año con mayor deficiencia en la información.

Gráfico 2. Distribución de la edad por periodo anual en las bases de datos disponibles



La edad corresponde a la diferencia entre la fecha de nacimiento y la fecha de prestación del servicio para cada servicio registrado. Para el cálculo se eliminaron los registros con fecha de nacimiento en blanco o con dato no interpretable ("\N").

A partir del análisis presentado se seleccionaron los datos para los dos abordajes del análisis así:

1. Análisis prospectivo para construcción de un modelo multivariante. Se seleccionaron sujetos con y sin diagnóstico de anemia en el año 2011 para asegurar un seguimiento de 5 años (Ver descripción de método de análisis en la sección 8.6). La selección del año 2011 se debe a que el año 2012 presentó mayores deficiencias que los demás años y la selección de sujetos en el año 2013 no permitía completar 5 años de seguimiento dada la disponibilidad de información hasta el año 2017.

2. Análisis retrospectivo para construcción de un modelo basado en aprendizaje de máquinas. Se seleccionaron sujetos con y sin diagnóstico de cáncer gástrico en 2017 para asegurar un periodo de observación de dos años (Ver descripción de método de análisis en la sección 8.7). El año 2017 corresponde al último año de información disponible y el periodo de dos años de observación se basa en los descrito sobre la historia natural del cáncer gástrico, en donde los estudios de seguimiento disponibles muestran que con un periodo de seguimiento de dos años se podría identificar hasta el 90% de los casos en estados tempranos (36).

Para las poblaciones seleccionadas los valores faltantes de edad y sexo fueron reconstruidos a partir de la identificación del sujeto (ID) en registros de otros periodos en donde estuvieran consignados los valores de estas variables.

7.5 Etiquetado de casos y controles

La valoración de la validez de la información sobre diagnósticos de cáncer se hizo mediante análisis del etiquetado. La codificación básica del diagnóstico corresponde a los códigos del CIE10 asignado a cada uno de los registros (atenciones); no obstante, para verificar su validez se realizaron tres procedimientos que incluyen:

1. Una aproximación exploratoria mediante la estimación global del número de registros utilizando 5 algoritmos de etiquetado diferentes.
2. Determinación de la sensibilidad y especificidad de la codificación del CIE10 registrada en las bases de datos de la UPC.
3. Comparación del número de casos de cáncer gástrico prevalentes estimados a partir de los datos del Registro Poblacional de Cáncer de Bucaramanga, con el número de casos prevalentes identificados en las BDUPC.

7.5.1 Análisis exploratorio del comportamiento de las etiquetas

Para la identificación de casos de cáncer gástrico se propusieron diferentes alternativas de etiquetado a saber:

1. Códigos CIE10 para cáncer gástrico
    - Cáncer invasor C160 a C169
    - Cáncer in situ D002
2. Códigos CUPS relacionados con cáncer gástrico mas códigos CIE10 relacionados con enfermedad benigna del estómago
    - Códigos CUPS cáncer gástrico
        - Gastrectomía parcial 43.8 y derivados
        - Gastrectomía total 43.9 y derivados
        - Gastrectomía parcial con reconstrucción 43.7 y derivados
        - Infusión de quimioterapia 99.2.5 y derivados
        - Administración de radioterapia 92.2 y derivados
    - Códigos CIE10 enfermedad benigna del estómago
        - Úlcera gástrica K250 a K259
        - Úlcera duodenal K260 a K269
        - Úlcera péptica no especificada K270 a K279

        - Úlcera gastro-yeyunal K280 a K289
        - Gastritis y duodenitis K290 a K299
        - Dispepsia funcional K30X
        - Otras enfermedades del estómago y del duodeno K310 a K319

3. Códigos CUPS para cáncer en general más códigos CIE10 para enfermedad benigna del estómago
    - Códigos CUPS para cáncer en general
        - Infusión de quimioterapia 99.2.5 y derivados
        - Administración de radioterapia 92.2 y derivados
    - Códigos CIE10 para enfermedad benigna del estómago como descritos

4. Código de medicamentos para cáncer gástrico más códigos CIE10 para enfermedad benigna del estómago.
    - Códigos ATC homologados correspondientes a los medicamentos definidos en las guías de práctica clínica para tratamiento de cáncer gástrico de NCCN vigentes a 2013 (81); esto debido a que para la fecha no se contaba con guías nacionales disponibles.
    - Códigos CIE10 para enfermedad benigna del estómago como descritos

5. Código de medicamentos para cáncer en general más códigos CIE10 para enfermedad benigna del estómago.
    - Códigos ATC homologados. Grupo L de la clasificación (Antineoplásicos e inmunomoduladores).

Debido al amplio uso de códigos CUPS en el análisis exploratorio y a los cambios introducidos en la codificación de procedimientos en el año 2015, se decidió utilizar la información del año 2013 para la validación de etiquetas, tomando en cuenta además que la base de datos de este periodo calendario contiene información sobre medicamentos como se describe en la Tabla 2.

En general, el uso de códigos de enfermedad benigna del estómago combinado con códigos CUPS y de medicamentos específicos de cáncer gástrico arrojó un número pequeño de registros, lo cual sugiere una relación desfavorable entre el beneficio y el elevado costo computacional (Tabla 4).

El número de registros se incrementó de manera significativa cunado se utilizaron los códigos CIE10 de enfermedad benigna del estómago combinados con códigos CUPS y medicamentos de cáncer en general (no específicos del cáncer de estómago), siendo particularmente llamativo el incremento en el número de registro con la combinación de códigos CUPS, la cual representa un aumento de 61,6% en relación con el número de registros cuando se utilizan únicamente los códigos CIE10 específicos de cáncer gástrico. En este caso, consideramos que el etiquetado puede prestarse a confusión toda vez que es posible que haya enfermedad ácido-péptica asociada al tratamiento del cáncer, la cual se expresa en síntomas y diagnósticos relacionados (CIE10 K290 a K319), sin que en realidad se esté hablando de un caso nuevo de cáncer gástrico. Esto puede ser particularmente cierto para pacientes que reciben quimioterapia o radioterapia por otros tipos de cáncer diferentes del cáncer gástrico. En esta medida, se consideró

que la utilización de códigos de CIE10 específicos de cáncer gástrico corresponde al etiquetado con más consistencia en el análisis exploratorio.

Tabla 4. Número de registros según tipo de etiquetado para el año 2013

Etiquetado	Total
CIE10 Cáncer gástrico	67253
CUPS Cáncer gástrico + CIE 10 Enfermedad benigna estómago	37
CUPS Cáncer (general) + CIE10 Enfermedad benigna estómago	108657
Medicamento cáncer gástrico + CIE10 Enfermedad benigna estómago	29
Medicamento cáncer (general) + CIE10 Enfermedad benigna estómago	455

7.5.2 Determinación de sensibilidad y especificidad de la codificación CIE10

En razón al desempeño en el análisis exploratorio, en este procedimiento se utilizó únicamente la codificación CIE10 para cáncer gástrico. Se tomó una muestra aleatoria de la base de datos del año 2013 y se utilizó como patrón de comparación el diagnóstico histopatológico consignado en la historia clínica del Hospital Universitario San Ignacio (HUSI).

Debido a que los datos se encuentran anonimizados (el identificador corresponde a un código encriptado), se desarrolló un algoritmo de emparejamiento para identificar los sujetos registrados en el HUSI. El algoritmo se desarrolló con datos del mes de julio de 2013 y una vez validado se realizó el emparejamiento para los restantes meses del año.

Los pasos para el desarrollo del algoritmo de emparejamiento comprenden:

1. Limpieza de datos y preparación de los conjuntos de datos. Para la base de datos del HUSI, se fusionaron las bases de datos de consulta de urgencias, consultas ambulatorias y egresos hospitalarios. Para la BDUPC se filtraron los registros de los servicios de salud prestados en el HUSI con base en el código de habilitación de dicho hospital. Dado que el HUSI informa datos de tres ubicaciones (Edificio principal, Centro de Cuidados Paliativos, Instituto de Memoria y Envejecimiento), solo se utilizaron datos del edificio principal (código HUSI 01).

   Como se observa en la Tabla 5, algunas variables tenían descripciones diferentes en ambas bases de datos, por lo que se unificaron códigos y nombres. En ambas bases de datos se filtraron los duplicados que tuvieran los mismos valores para el ID, sexo, fecha de nacimiento, fecha de prestación del servicio de salud, código CIE10 y código CUPS. También se eliminaron los registros con datos ausentes en una o más variables de interés.

2. Desarrollo de los algoritmos de emparejamiento. Para comparar los registros se utilizó un enfoque determinístico y uno probabilístico. Para el enfoque determinístico, analizar todas las combinaciones posibles entre registros BDUPC y HUSI tendría un alto costo computacional ya que el conjunto resultante sería N-UPC*N-HUSI = 5.214.343.551. Por ello, inicialmente se buscaron registros con 100% de coincidencia en fecha de nacimiento, sexo y fecha de prestación del

servicio de salud. Dado que una persona puede tener varios registros en un día (todos los servicios prestados ese día), se filtraron los duplicados por ID.

Tabla 5. Disponibilidad y utilización de variables para el emparejamiento de registros

Variable	BDUPC	HUSI
Utilizadas para emparejamiento		
ID	No	Si
Aseguradora	Código	Nombre
Fecha de nacimiento	Si	Edad
Sexo	Si	Si
Fecha de procedimiento	Si	Si
Código de procedimiento (CUPS)	Si	Si
Código del centro de salud	Si	Si
Código de diagnóstico (CIE10)	Si	Si
Otras variables		
Municipio	Código	Nombre
Tipo de cuidado (Hospital/Ambulatorio)	Si	No
Días de estancia	Si	No
Código de medicamento	No	Si
Nombre de medicamento	Si	Si

Luego de obtener el conjunto de datos mediante el enfoque determinístico, se calculó la probabilidad de que dos registros fueran iguales comparando dos variables en cada par: aseguradora y procedimiento (código CUPS). Las cadenas de texto para ambas variables se compararon con los algoritmos de Jaro, Jaro-Winkler y Damerau-Levenshtein, que son medidas métricas de cadenas (82, 83). Jaro considera la cantidad de caracteres coincidentes, la cantidad de transposiciones y la longitud de las cadenas; Jaro-Winkler añade a la medida de Jaro una escala de prefijos que da una respuesta más precisa cuando las cadenas tienen un prefijo común dentro de una longitud máxima. Damerau-Levenshtein considera el número mínimo de operaciones (inserciones, eliminaciones, sustituciones y transposiciones) necesarias para cambiar una palabra por otra. El valor de la comparación osciló entre 0 y 1 (probabilidad de coincidencia), en donde 0 significa cadenas diferentes y 1 significa cadenas idénticas.

La matriz de probabilidad obtenida previamente se utilizó como dato de entrada para la vinculación de registros. Dado que no hay una base de datos que constituya el estándar de oro, se utilizó el algoritmo ECM no supervisado, el cual es un algoritmo probabilístico iterativo (84). Para el análisis de datos se utilizaron las bibliotecas Record Linkage y Pandas en Python®.

3. Validación de los algoritmos de emparejamiento. Se realizó mediante 1) la comparación de las características básicas de registros emparejados y no emparejados y 2) el análisis de sensibilidad con una variable no incluida en el algoritmo de emparejamiento (85). Se compararon registros emparejados y no

emparejados con base en los resultados del algoritmo Jaro-Winkler (acuerdo total para aseguradora y procedimiento vs. acuerdo parcial o no acuerdo), utilizando las variables sexo, edad (promedio) y lugar de residencia (Bogotá vs. otro).

El análisis de sensibilidad se realizó utilizando la variable municipio, la cual no fue incluida en los procedimientos de emparejamiento. Estimamos un tamaño de muestra de 283 registros para ser analizados con una Kappa mínimo aceptable de 0,75 (K0), Kappa esperado 0,9 (K1), proporción desconocida de resultados (p: 0,5), nivel de significancia 0,05 (alfa), potencia 80% (beta), y tasa de deserción 0. Se obtuvo una selección aleatoria de registros sin duplicados y con coincidencia completa con el algoritmo Jaro-Winkler.

Finalmente, se analizaron diferentes puntos de corte comparando los algoritmos de Jaro-Winkler y ECM (81, 83), utilizando la muestra seleccionada para la validación de municipio y definiendo como verdaderas parejas aquellos registros con el mismo municipio y falsas parejas aquellos registros con lugares de residencia diferentes en las dos bases d datos.

La probabilidad promedio de emparejamiento para las dos variables utilizadas fue:

- Aseguradora: 0,99 para los algoritmos Jaro-Winkler y Jaro y 0,98 para el algoritmo Damerau-Levenshtein.
- Servicio/procedimiento: 0,60 para Jaro–Winkler, 0,59 para Jaro y 0,31 para Damerau–Levenshtein.

Luego de filtrar duplicados y eliminar registros con variables sin datos, obtuvimos 124.599 registros en la base de datos HUSI y 41.849 registros en la base de datos de la UPC. Luego del algoritmo determinístico encontramos 477.724 coincidencias, y luego de filtrar duplicados por código de identificación, se obtuvieron 8.995 individuos con emparejamiento total o parcial (Tabla 6).

Tabla 6. Emparejamiento total y parcial de registros

Sex	Pareja A-P	Pareja solo A	Pareja solo P	Sin pareja
	(Jaro-Winkler 1-1)	(Jaro-Winkler 1-0)	(Jaro-Winkler 0-1)	(Jaro-Winkler 0-0)
Men	1168	2326	83	156
Women	1867	3036	129	175
Total	3035	5362	212	331

A: Aseguradora. P: procedimiento médico

Con base en estos resultados, se consideró que se puede utilizar el algoritmo de Jaro-Winkler si la probabilidad de coincidencia es de 0,7 o mayor (punto de corte).

La comparación de registros emparejados y no emparejados mostró que la edad y el lugar de residencia no tuvieron diferencias significativas, mientras que las mujeres fueron un 6% más en los registros coincidentes (Tabla 7).

En el análisis de sensibilidad mediante la verificación del lugar de residencia (municipio) entre registros emparejados con el algoritmo de Jaro-Winkler (1-1), encontramos una concordancia superior al 90% con índice Kappa entre 0,6 y 0,8 ($p < 0,0001$), dependiendo de la interpretación de datos, de la siguiente manera:

• Concordancia con datos sin filtrar: Kappa 0,6.

• Concordancia asumiendo el municipio como lugar de nacimiento o residencia: Kappa 0,7.

• Concordancia excluyendo registros que solo tenían una única atención de urgencias en el HUSI: Kappa 0,8

Tabla 7. Comparación de registros emparejados y no emparejados

Variable	Registros emparejados	Registros no emparejados
Edad (promedio)	38,6	37,9
Sexo		
Mujeres	1.168 (38,5%)	2.565 (43.4%)
Hombres	1.867 (61.5%)	3.340 (56.6%)
Residencia		
Bogotá	2.619 (86.3%)	5.361 (90.8%)
Otra	416 (13.7%)	544 (9.2%)
Total	3.035	5.905

La concordancia fue significativamente mayor para los registros con Bogotá como lugar de residencia que para otros municipios, pero asumiendo la concordancia para los datos sin filtrar, no encontramos diferencias significativas entre los algoritmos de Jaro-Winkler y ECM para un punto de corte de 0,7 o mayor (probabilidad de coincidencia).

Una vez emparejados los registros se obtuvo una muestra aleatoria de casos etiquetados como cáncer de estómago durante todo el año 2013. La sumatoria de registros mensuales en el HUSI para el año 2013 fue 582.826 (promedio mensual 48.568,8), con sumatoria de registros únicos 196.370 (promedio mensual 16.364,2). Para una sensibilidad esperada de 70% y precisión de 10% se estimó una muestra de 81 registros de cáncer gástrico a revisar frente a la historia clínica del HUSI. Con una muestra de 20.000 registros emparejados (sumatoria de registros únicos por mes) se obtuvieron 66 registros con código CIE10 correspondiente a cáncer gástrico en la base BDUPC (C160 a C169 y D002) y 83 registros de cáncer gástrico en la base de datos HUSI. El cálculo de medidas de desempeño se hizo por el cruce de códigos y solo se revisaron frente a la historia clínica del HUSI los sujetos identificados con diagnóstico de cáncer gástrico en las bases de datos (92 en total); es decir, los sujetos sin reporte de cáncer gástrico en ambas bases de datos se asumieron como verdaderos negativos. La sensibilidad obtenida fue 68,7% (IC95% 57,6%-78,4%) y la especificidad 99,9% (IC95% 99,91%-99,98%). Los valores para el cálculo se presentan en la tabla de 2 por 2.

		HUSI		Total
		C16	No C16	
BDUPC	C16	57	9	66
	No C16	26	20153	20179
Total		83	20162	20244

La elevada especificidad observada y el elevado valor predictivo positivo (86,4%) permite concluir que hay una baja probabilidad de que se asignen los códigos CIE10 de cáncer gástrico a sujetos sin este diagnóstico; sin embargo, el hallazgo de 31,3% de falsos negativos (1-sensibilidad) sugiere nuevamente la necesidad de utilizar un complemento del etiquetado mediante códigos de cáncer gástrico en el CIE10 si la búsqueda de casos arroja un valor menor al esperado.

7.5.3 Estimación del número de casos de cáncer gástrico prevalentes en Bucaramanga

El Registro Poblacional de Cáncer de Bucaramanga se encuentra validado por el sistema global de vigilancia de la incidencia de cáncer de la Agencia Internacional de Investigación en Cáncer (IARC por su sigla en inglés) y la OMS (86).

Sus métodos implican la recolección sistemática de los casos nuevos de cáncer en todas las posibles fuentes de información de Bucaramanga y su área metropolitana incluyendo egresos hospitalarios, consultas hospitalarias y de centros de tratamiento oncológico, consulta privada de especialistas en áreas de oncología, laboratorios de patología, centros de radiodiagnóstico y certificados de defunción. Los datos recolectados sufren un estricto proceso de validación que elimina duplicados, elimina casos prevalentes (diagnosticados en periodos previos) e identifica segundos tumores primarios en un mismo sujeto (caso nuevo de cáncer en un sujeto con otro tipo de cáncer).

La determinación del número de casos nuevos por año con elevada precisión permite la estimación de prevalencias mediante el método definido por la IARC (87):

$$P(n - casos\ año) = \sum_{i=1}^{n} ICx * S(i - 0{,}5)$$

En donde IC_x es el número anual de casos nuevos a la edad X; *S* es la proporción de casos diagnosticados a la edad X que siguen vivos tras el tiempo *t* transcurrido desde el diagnóstico; y *n* es el número de años como caso, o lo que es igual, el número de años hasta ser considerado curado. La definición de curación en cáncer es controversial, pero en general, el seguimiento estricto se propone por un periodo de 5 años, luego de los cuales se considera que el paciente está libre de enfermedad si esta ha remitido completamente. En consecuencia, la estimación de supervivencia a 5 años es considerada como el desenlace principal en el manejo de la enfermedad (88).

Para estimar los casos prevalentes en Bucaramanga se asumieron casos diagnosticados en todas las edades y se partió de la supervivencia a 5 años estimada para cáncer de estómago en Colombia según datos del Instituto Nacional de Cancerología (24,6%)(89), y del programa de vigilancia global de la supervivencia de cáncer de la London School of Hygiene and Tropical Medicine (Programa CONCORD: 17,1%)(4).

El número de casos prevalentes en la BDUPC se determinó mediante dos definiciones de las etiquetas a saber: solo registros con códigos CIE10 para cáncer gástrico (mayor consistencia en el análisis exploratorio) y la suma de registros con código CIE10 para cáncer gástrico y registros con CUPS de gastrectomía independiente del CIE10 asociado en búsqueda de mayor sensibilidad. Los códigos CUPS para gastrectomía se mantienen sin variación para los años 2013 (Resolución 1896 de 2001) y 2017 (Resolución 4678 de 2015 y Resolución 1132 de 2017).

Tabla 8. Estimación de casos de cáncer gástrico prevalentes en Bucaramanga

Año	Casos incidentes RPC Bucaramanga	Casos prevalentes Supervivencia INC		Casos prevalentes Supervivencia CONCORD		Casos prevalentes BDUPC - Etiqueta 1	Casos prevalentes BDUPC - Etiqueta 2
Prevalencia 2013							
2009	152	151		117			
2010	157	131		94			
2011	163	97		70			
2012	168	60		43			
2013	194	21	**460**	17	**341**	**213**	**284**
Prevalencia 2017							
2013	194	186		149			
2014	205	167		123			
2015	194	126		83			
2016	230	72		59			
2017	239	28	**579**	20	**434**	**79**	**282**

RPC: Registro Poblacional de Cáncer. La etiqueta 1 de la BDUPC corresponde a códigos CIE10 para cáncer gástrico y la etiqueta 2 corresponde a códigos CIE10 para cáncer gástrico mas códigos CUPS de cirugía asociada a cáncer gástrico (gastrectomía) independiente del código CIE10 asociado.

La interpretación de resultados de la Tabla 8 debe tomar en cuenta que la estimación de prevalencia es para el total de la población del área metropolitana de Bucaramanga (incluye municipios aledaños Floridablanca, Girón, Piedecuesta), mientras que las estimaciones de la BDUPC corresponden únicamente a la población de régimen contributivo en Bucaramanga. La cobertura de aseguramiento del régimen contributivo para los años 2013 y 2017 en el área metropolitana de Bucaramanga se encontraba en 67,8% Y 67,4% respectivamente (75). Utilizando la misma distribución porcentual el número de casos prevalentes esperados para la población de régimen contributivo sería 311 y 231 en 2013 según las estimaciones del INC y el programa CONCORD respectivamente; y 390 y 293 para el año 2017.

Según estos resultados, el etiquetado 2 de la BDUPC muestra mayor consistencia y presenta valores más cercanos a la estimación de prevalencia, particularmente cuando se utiliza la estimación de supervivencia del programa CONCORD, con lo cual además se identifica un número mayor de casos sugiriendo una reducción en el número de falsos negativos (mayor sensibilidad). No obstante, como se anotó el costo computacional de la combinación es alto e incrementa la probabilidad de falsos positivos, de forma que el uso de los códigos de gastrectomía debe valorarse en relación con la necesidad de identificar más casos si la muestra obtenida no tiene tamaño suficiente, o frente a la necesidad de eliminar falsos negativos en el grupo de controles a fin de mejorar la precisión de los modelos predictivos.

7.6 Construcción de un modelo multivariado de predicción clínica

La Figura 2 resume la construcción del modelo multivariado de predicción clínica en cáncer gástrico.

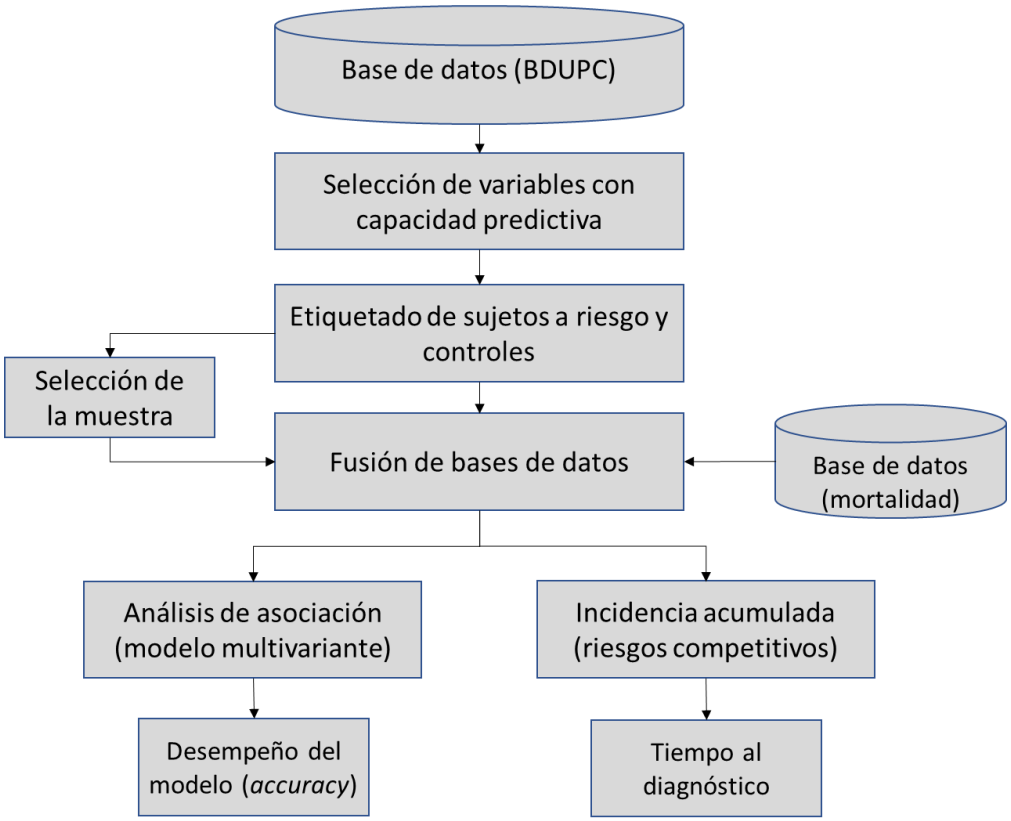


Figura 2. Flujograma para la construcción del modelo multivariante

7.6.1 Análisis exploratorio para la identificación de variables predictoras

Se utilizó LASSO para identificar potenciales variables predictoras del diagnóstico de cáncer gástrico (90). Se realizó un análisis paramétrico para la selección de la constante de penalización (lambda) y una validación cruzada de 10 por 10. Los parámetros finales para el modelo fueron número máximo de iteraciones 200, valor de t 1e-10, valor de lambda 0,005 y razón de entrenamiento 0,5. El parámetro de penalización o regularización lambda limita el tamaño de los coeficientes, llevando algunos de ellos a valor cero con el finde eliminar variables que no tienen capacidad predictiva (67).

En razón a que el seguimiento longitudinal se definió a partir de la selección de la muestra en 2011 (Ver sección 8.4), el análisis exploratorio se hizo seleccionando casos y controles con datos de septiembre a diciembre de 2013, ya que el año 2012 mostró mayores deficiencias en la calidad de la información.

Se utilizó un periodo de seguimiento de 12 meses previo al diagnóstico; es decir desde octubre de 2012. En este caso los datos de 2012 utilizados corresponden a pacientes claramente caracterizados en el último trimestre de 2013; no obstante, la base de datos de 2012 no tiene información de medicamentos por lo que los meses de ese año incluidos en el análisis no aportan datos a dicha variable.

En total 8.234.601 pacientes utilizaron servicios de salud de septiembre a diciembre de 2013, con un total de 41.152.616 registros. De esta población 305 pacientes tuvieron diagnóstico de cáncer gástrico, definido como la coincidencia del código CIE10 para

cáncer gástrico (C160 a C169 y D002) y el código CUPS para el estudio de biopsia (89.8.1.01); es decir, estos corresponden a diagnóstico nuevo de cáncer. La distribución de casos y controles se muestra en la Tabla 6, en donde se observa que hay una relación inversa en la distribución de la edad y el sexo entre casos y controles.

Tabla 9. Distribución por edad y sexo de la muestra para el análisis exploratorio de variables predictoras en la construcción del modelo multivariante

Variable	Cáncer gástrico	Controles
Edad		
<30	2 (0,7)	3.275.995 (39,8)
30-49	53 (17,4)	2.301.198 (27,9)
50-60	70 (23,0)	1.158.101 (14,1)
>60	180 (59,0)	1.499.002 (18,2)
Sexo		
Hombre	174 (57,0)	3.296.309 (40,0)
Mujer	131 (43,0)	4.937.987 (60,0)

Los porcentajes corresponden al total de la variable en las columnas

Durante el período de seguimiento se identificaron 1.809 diagnósticos y en total 5.172 procedimientos y 703 medicamentos fueron utilizados por los pacientes de la muestra. Tras la revisión de los resultados iniciales se eliminaron diagnósticos, procedimientos y medicamentos relacionados con el tratamiento del cáncer gástrico debido a que la búsqueda de variables predictoras debe ser previa al diagnóstico. Igualmente, se eliminaron diagnósticos relacionados con trauma (causas externas), síntomas mal definidos y procedimientos de rutina (ej. consulta de control).

Finalmente quedaron 59 diagnósticos (agrupación de 3 dígitos del CIE10), 108 procedimientos y 62 medicamentos a ser incluidos en el modelo. Adicionalmente se utilizó la información de edad (mayor o menor de 40 años), sexo y municipio (80 valores). Para el análisis cada diagnóstico, procedimiento, medicamento y municipio constituyen una variable binaria (0-1), lo que resultó en 311 variables en el modelo.

Tabla 10. Número de variables por categoría que mostraron asociación significativa con el diagnóstico de cáncer gástrico en el análisis exploratorio para la construcción del modelo multivariante

Variable	Asociación positiva	Asociación negativa	Total
Diagnóstico	8	2	10
Procedimiento	37	3	40
Medicamento	11	10	21
Edad	0	2	2
Sexo	0	0	0
Municipio	5	7	12

En total 85 variables mostraron asociación significativa con el diagnóstico de cáncer de estómago (coeficiente diferente de 0), 61 de ellas con asociación positiva y 24 con asociación negativa (Tabla 10).

Los diagnósticos y procedimientos relacionados corresponden a anemias carenciales, infección, malnutrición, y eventos trombóticos; siendo la anemia la que mayor OR reportó. La caracterización de los diagnósticos y los procedimientos potencialmente asociados se presentan en la Tabla 11.

Las anemias carenciales se identifican como la variable con mayor valor predictivo. La anemia por deficiencia de folatos (código D52 en CIE10) reporta OR 6,77 (IC95% 1,0-12,7) y a pesar de que el intervalo de confianza no indica significancia estadística, se observan procedimientos asociados a otras anemias carenciales como el suministro de cianocobalamina o vitamina B12 asociado con anemia perniciosa (código D51 en CIE10), y la determinación de ferritina y hierro total relacionada con anemia ferropénica (código D50 en CIE10). Algunas infecciones aparecen también como potencial predictor, lo cual es coherente con la asociación de inmunosupresión y riesgo incrementado de cáncer gástrico descrita previamente (91); sin embargo, la extensa y poco específica codificación de CIE10 para condiciones infecciosas (códigos A y B más códigos específicos por localización anatómica de la infección), limitan la posibilidad de usar esta codificación como variable predictiva en un modelo multivariante. Se observaron otros potenciales predictores, pero con menor fuerza de asociación (Tabla 11).

Con base en los hallazgos descritos, se decide utilizar las anemias carenciales como la variable de exposición. En razón a la asociación potencial de otros códigos diagnósticos se propone utilizar el índice de comorbilidad de Charlson (92), el cual incluye trastornos cardiovasculares, condiciones asociadas a inmunosupresión, demencia y diagnósticos diferenciales del cáncer gástrico (Ver sección 8.6.3).

7.6.2 Etiquetado de sujetos a riesgo y controles

Se hizo un análisis de sensibilidad del etiquetado de anemia con la información de julio de 2013, utilizada en el proceso de validación del emparejamiento de registros entre la BDUPC y la base de datos del HUSI. Se usó una probabilidad de coincidencia ≥ 0,7 para la selección de parejas (Ver sección 8.5.2). Se identificaron 6.748 registros emparejados que reportan 16 casos de anemia según los códigos descritos.

La comparación del diagnóstico de anemia en la BDUPC con los registros clínicos del HUSI como patrón de oro (diagnóstico o antecedente de anemia registrado), arrojó una sensibilidad de 54,5% (IC95% 23,4%-83,2%) y especificidad de 99,9% (IC95% 99,8%-99,9%), lo que indicaría una baja proporción de falsos positivos. Sin embargo, en este caso el valor predictivo positivo fue 54,1%, dato afectado por la prevalencia de enfermedad en la muestra (0,16%), la cual podría ser más alta que en la población general dada la condición de hospital de alta complejidad del HUSI. El valor predictivo observado equivale a que el 45,9% de registros con diagnóstico de anemia en la BDUPC no tenían ese diagnóstico en la historia clínica del HUSI; tomando en cuenta que para la selección de sujetos a riesgo es más relevante una baja proporción de falsos positivos y que el número de casos en la base total del país es suficientemente alto, se decide seleccionar los casos con diagnóstico de anemia solo si este aparece en dos o más meses diferentes del año 2011.

Tabla 11. Diagnósticos asociados positivamente al diagnóstico de cáncer gástrico y procedimientos y medicamentos relacionados con los diagnósticos en el análisis exploratorio para la construcción del modelo multivariante

Variable	Tipo de variable	OR	IC95%	
			LI	LS
Enfermedades potencialmente asociadas				
Anemia				
Anemia por deficiencia de folatos	Diagnóstico	6,77	1,00	12,74
Cianocobalamina	Medicamento	2,09	1,44	2,81
Vitamina B12	Medicamento	1,11	1,00	1,60
Hemograma*	Procedimiento	2,44	1,96	3,14
Prueba de compatibilidad cruzada	Procedimiento	1,24	1,00	1,71
Ferritina	Procedimiento	1,19	1,00	1,79
Hierro total	Procedimiento	1,18	1,00	1,62
Infección				
Mastoiditis	Diagnóstico	3,01	1,00	10,24
Tuberculosis respiratoria	Diagnóstico	1,31	1,00	3,10
Radiografía de tórax*	Procedimiento	1,31	1,01	1,64
Drenaje de colección	Procedimiento	1,03	1,00	1,30
Trastornos nutricionales				
Proteínas totales	Procedimiento	1,65	1,05	2,60
Albúmina	Procedimiento	1,62	1,29	2,04
Consulta nutrición	Procedimiento	1,09	1,00	1,46
Trastornos de coagulación				
Tiempo de protrombina	Procedimiento	1,27	1,01	1,59
Recuento de plaquetas	Procedimiento	1,03	1,00	1,29
Otras enfermedades (no se descarta comorbilidad)				
Demencia				
Enfermedad de Alzheimer	Diagnóstico	1,40	1,00	2,37
Encefalograma	Procedimiento	1,19	1,00	1,94
Hiperplasia prostática	Diagnóstico	1,01	1,00	1,05
Afecciones cardiovasculares				
Ecocardiograma	Procedimiento	1,05	1,00	1,18
Perfusión miocárdica	Procedimiento	1,02	1,00	1,15
Nifedipina	Medicamento			
Enalapril	Medicamento	1,02	1,00	1,16
Enfermedad de cuello uterino y vagina				
Trastornos no inflamatorios del cérvix	Diagnóstico	1,11	1,00	1,88
Trastornos del sueño	Diagnóstico	1,01	1,00	1,17
Hipoacusia				
Otras hipoacusias	Diagnóstico	1,02	1,00	1,20
Audiometría	Procedimiento	1,08	1,00	1,38

** El hemograma y la radiografía de tórax pueden estar relacionados con múltiples diagnósticos de los que mostraron asociación*

De otra parte, en la verificación de diagnósticos de anemia en los registros clínicos se observó dificultad en precisar el tipo de anemia, por lo que se decide utilizar todos los códigos de anemia excluyendo las anemias hemolíticas, por considerar que estas tienen vía etiológica particularmente diferente. En conclusión, se utilizaron los códigos D50 a D53 y D60 a D64.

7.6.3 Selección de la muestra y fusión de las bases de datos

Para la selección de la muestra se incluyeron sujetos con edad entre 40 y 90 años que recibieron alguna atención en el año 2011. Se excluyeron pacientes con diagnóstico de cáncer de cualquier tipo (CIE10 C00 a C97 y D00 a D09) y pacientes reportados en la base de datos de mortalidad para el año 2011.

Una vez filtrados los registros por los criterios de elegibilidad, se clasificaron los pacientes según la condición “con anemia” y “sin anemia” durante el año 2011. Anemia se definió según el etiquetado descrito en la sección 8.6.2. La condición “sin anemia” se definió como la ausencia total de registros con CIE10 codificados como anemia durante 2011 (D50 a D64). Se construyeron bases de datos independientes para los pacientes “con anemia” y “sin anemia”, y mediante un proceso iterativo y la revisión individual de registros, se depuró la base de datos (ej. sujetos “con anemia” pero con registro un solo mes, sujetos incluidos en las dos bases de datos, etc).

Tras depurar los registros se eliminaron los duplicados por ID, lo que arrojó 14.399 individuos con anemia y 44.961 individuos sin anemia. Con seguimiento a 5 años la muestra seleccionada aportaría hasta 71.995 y 224.805 años persona respectivamente. Con una prevalencia esperada de cáncer gástrico de 0,06% en los sujetos sin anemia (60 por 100.000 estimado global para Colombia en el rango de edad)(22), un RR de 1,8 (21), confianza 95% (alfa 0,05) y poder de 80% (beta 0,2), se necesitarían 57.289 años persona en cada grupo para evaluar la asociación entre la condición “con anemia” y el diagnóstico de cáncer gástrico (93), por lo que se asume que tras los filtros aplicados el tamaño de muestra obtenido es adecuado para el análisis.

Una vez seleccionados los sujetos de estudio, se fusionaron las bases de datos de los años 2012 a 2016 para identificar todas las atenciones que recibió en este periodo la población de estudio. Inicialmente se construyó una base de datos con formato *long* en donde cada atención de un mismo paciente se incluyó como registro independiente.

Posteriormente se identificaron pacientes de la muestra seleccionada que estaban incluidos en la base de datos de mortalidad. Para los pacientes fallecidos el registro de mortalidad se incorporó como un registro de atención dentro del formato *long* con el ID del paciente. Las variables tomadas de la base de datos de mortalidad fueron ID, causas de muerte (todas las registradas en la secuencia de eventos del certificado de defunción), fecha de defunción (todas las registradas para los eventos que llevaron a la muerte).

Tras la fusión de la base de datos de mortalidad se identificaron sujetos que reportaban diagnóstico de cáncer posterior a la fecha de muerte. El análisis de la secuencia lógica de eventos y la revisión de fechas en el registro correspondiente a la defunción permitió corregir la información en la mayoría de casos, solo se excluyeron los registros de 3 pacientes (excluidos por ID), debido a que los datos no permitieron establecer una secuencia lógica de eventos.

Tabla 12. Estructura de las bases de datos para la construcción del modelo multivariante

Variables fusión 2011-2016 (*long*)	Variables fusión mortalidad (*long*)	Variables base de datos final (*wide*)
Variables de estudio ID Aseguradora[1] Fecha de nacimiento Sexo Municipio[3] Diagnóstico[5] Fecha de primera atención Procedimiento (CUPS) Fecha de última atención Medicamento[7] Otras variables Ámbito del procedimiento Forma de reconocimiento Días de estancia Tipo de servicio Valor pagado Código IPS	Variables de estudio ID Aseguradora[1] Fecha de nacimiento Sexo Municipio[3] Diagnóstico[5] Fecha de primera atención Procedimiento (CUPS) Fecha de última atención Medicamento[7] Muerte[4] Fecha de muerte Causa de muerte (CIE10)[9]	Variables de estudio ID Edad[2] Sexo Municipio[3] Municipio capital[4] Diagnóstico de cáncer gástrico[4] Diagnóstico de anemia en 2011[4] Diagnóstico de anemia 2012-2016[6] Diagnóstico comorbilidad[8] Muerte[4] Muerte por cáncer gástrico[4] Fecha diagnóstico cáncer gástrico Fecha diagnóstico anemia 2012-16 Fecha de último seguimiento Fecha de muerte

1. Código EPS; 3. Fecha de primera atención menos fecha de nacimiento; 3. Código DIVIPOLA solo para el municipio de diagnóstico de cáncer gástrico o municipio de último seguimiento; 4. Si/No; 5. Código CIE10; 6. Solo para sujetos sin anemia en 2011; 7. Nombre, código CUMS o código ATC; 8. Se incluyen diagnósticos comprendidos en el índice de comorbilidad de Charlson; 9. Cáncer gástrico/Otra causa

Se realizó recodificación de variables para propósito del análisis (Tabla 12). La edad se estimó a la fecha de la primera atención durante 2011. La información de registros asociados a diagnósticos diferentes de cáncer gástrico o anemia según las etiquetas definidas se limitó a las causas de enfermedad incluidas en el índice de comorbilidad de Charlson. Se asumió como muerte por cáncer gástrico todos los registros en donde se registró este diagnóstico dentro de las causas de muerte independientemente de su posición en la secuencia de eventos, las restantes causas de muerte se codificaron como muerte por otra causa. Para los sujetos “sin anemia” durante el año 2011 se identificó si habían sido diagnosticados “con anemia” en los años subsiguientes, utilizando la etiqueta definida. Una vez recodificadas las variables se construyó una base de datos en formato *wide* con un único registro para cada sujeto de investigación y se procedió a fusionar las bases de datos de los sujetos “con anemia” y “sin anemia”.

El índice de comorbilidad de Charlson (92) se modificó para efectos del estudio; para su construcción se utilizaron los códigos descritos en la Tabla 13. Se excluyó el cáncer gástrico toda vez que es la variable de desenlace. El puntaje corresponde a la suma simple de los puntajes asignados a cada diagnóstico; es decir, no se toma en cuenta la edad ya que esta se manejó como variable independiente dentro del modelo predictivo. El índice es entonces una variable continua con rango de 0 a 29.

Tabla 13. Comorbilidad considerada en el análisis del modelo multivariante y codificación para construcción del índice de comorbilidad de Charlson

Diagnóstico	Puntaje	Código CIE 10
Infarto de miocardio	1	I21-I23
Disnea paroxística nocturna	1	R06
Enfermedad vascular periférica	1	I73
ACV o ICT	1	I60-I64
Demencia	1	F00-F03
EPOC	1	J40-J44
Enfermedad del tejido conectivo	1	M30-M36
Enfermedad úlcero-péptica	1	K20-K23 + K25-K30
Enfermedad hepática mayor: con hipertensión portal[1]	3	K70.3, K74 con K76.6
Enfermedad hepática menor: sin hipertensión portal	1	K73-K74
Diabetes mellitus[2]	1	E10.0, E10.1, E10.6, E11-E14
Diabetes con daño de órgano	2	E10.2, E10.3, E10.4, E10.5
Hemiplejía	2	G81
Enfermedad renal crónica (insuficiencia)	2	N18
Tumor sólido (excluye cáncer gástrico)[3]	2	C00-C80
Leucemia o linfoma	2	C81-C85 + C91-C95
SIDA	6	B20B-24

1. K76.6 debe estar presente; K70.3 y K76.6 deben incluir el cuarto dígito después del punto; 2. Solo si incluyen el cuarto dígito después del punto 3. Debe excluir cáncer gástrico (C16)

7.6.4 Análisis de asociación y riesgos competitivos

Se hizo una aproximación descriptiva de las variables. Se calculó la frecuencia de atenciones por año para pacientes con diagnóstico de cáncer gástrico, pacientes con anemia y pacientes sin anemia. Se analizó la tendencia de la frecuencia de atenciones en los pacientes con diagnóstico de cáncer gástrico asumiendo la fecha de diagnóstico de cáncer gástrico como la fecha de la primera atención reportada con este diagnóstico.

Se analizó la asociación del diagnóstico de cáncer con las variables predictoras disponibles incluyendo edad, sexo, anemia (2011), e índice de comorbilidad. El municipio no se consideró dada la cantidad de valores posibles (1087) y la limitación del dato solo para el diagnóstico de cáncer gástrico. Se calculó el riesgo relativo con su respectivo intervalo de confianza utilizando dos desenlaces: diagnóstico de cáncer gástrico o diagnóstico/muerte por cáncer gástrico. Se evaluó el modelo utilizando el diagnóstico de anemia posterior a 2011 como variable de ajuste (anemia 2012-2016), pero no fue posible estimar el parámetro ya que la variable solo se mide en el grupo de pacientes “sin anemia” (colinealidad). Se ajustó el riesgo relativo mediante un modelo de regresión logística, se siguió un proceso iterativo y se utilizó el criterio de información de Akaike (AIC) para la selección final del modelo. La capacidad predictiva del modelo se evaluó mediante el cálculo de sensibilidad, especificidad y área bajo la curva.

Adicionalmente se analizó el riesgo de diagnóstico de cáncer gástrico utilizando una función de incidencia acumulada y la prueba de Gray para analizar las diferencias (94). Se utilizó además un modelo de riesgos competitivos; los modelos de riesgos

competitivos son modelos multivariantes de tiempo al evento en los que se asume que cada individuo solo puede tener uno de los dos eventos en competencia y por tanto la ocurrencia de un evento diferente al evento de interés afecta la probabilidad de ocurrencia de este último. Para el presente estudio se asume que la muerte por una causa diferente del cáncer gástrico es un evento competitivo frente al diagnóstico de cáncer gástrico. Se utilizó la función de incidencia acumulada para un tiempo de seguimiento de 60 meses. El ajuste por riesgos competitivos se hizo mediante un modelo multivariado según Fine y Gray (94).

7.7 Construcción de un modelo predictivo basado en aprendizaje de máquinas

Debido a que en cáncer de páncreas solo se logró un análisis exploratorio para el año 2013, la información relativa a este ejercicio se presenta en el Anexo 6.

El proceso para la construcción del modelo en cáncer gástrico se resume en la Figura 3.

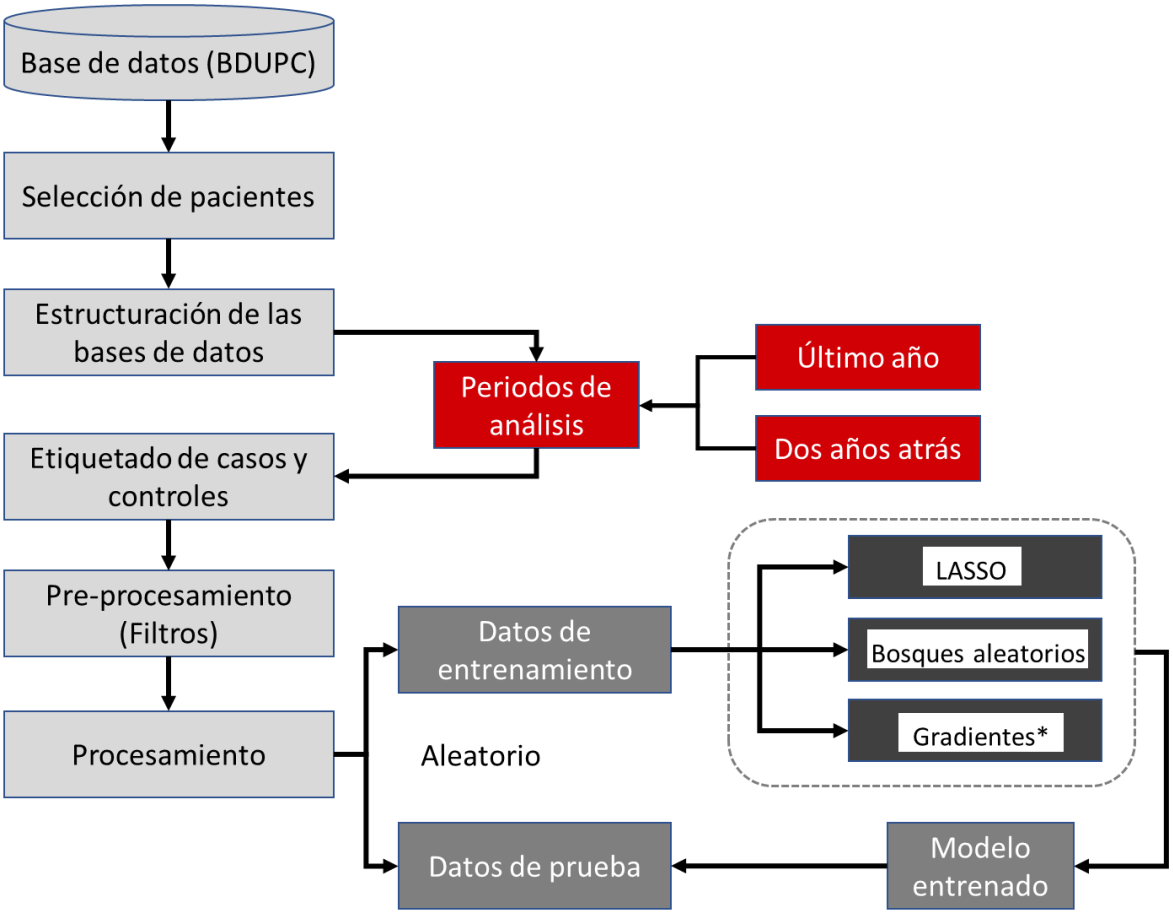


Figura 3. Flujograma para la construcción del modelo de aprendizaje de máquinas
**Potenciación de gradientes*

7.7.1 Selección de pacientes y estructuración de las bases de datos

La selección de pacientes se hizo para el periodo septiembre 1 a diciembre 31 de 2017 (Ver sección 8.4). Se eliminaron los registros sin dato de fecha de nacimiento en este periodo, se eliminaron los duplicados por ID y se seleccionó la población de 30 a 90 años.

Una vez seleccionada la población se estructuraron dos bases de datos correspondientes a diferentes periodos de análisis. Para analizar la capacidad predictiva del modelo en un periodo próximo al desenlace (diagnóstico de cáncer gástrico), se utilizaron los datos del año calendario en el que se seleccionó la población de estudio, esto es enero a agosto de 2017. Un segundo periodo de análisis corresponde a los dos años previos al periodo en donde se determina el desenlace (Ver sección 8.4), esto es de septiembre 1 de 2015 a agosto 31 de 2017. Sin embargo, a fin de reducir el efecto de diagnósticos y procedimientos cercanos al desenlace cuando la sintomatología es más evidente, se ignoró la información de los seis meses previos a la valoración del desenlace; es decir que para este periodo de análisis se tomaron registros de la población de estudio de septiembre 1 de 2015 a febrero 28 de 2017. Esto permite además reducir la colinealidad para la comparación de la capacidad predictiva del modelo en los dos periodos de análisis.

Las variables incluidas en el análisis son el ID, el sexo, la edad, la aseguradora (código EPS), el municipio (DIVIPOLA), el diagnóstico (código CIE10), y el procedimiento (código CUPS). La edad se calculó para los casos como la diferencia en años entre la fecha de diagnóstico de cáncer y la fecha de nacimiento, y para los controles como la diferencia en años entre la fecha de primera atención en el periodo de selección de la muestra (septiembre-diciembre 2017) y la fecha de nacimiento.

A fin de mejorar la eficiencia del procesamiento, los duplicados por ID, sexo y diagnóstico/procedimiento fueron eliminados; es decir, si en el periodo de estudio existían dos o más registros coincidentes en las tres variables se dejó uno solo de ellos. Adicionalmente, se excluyeron los pacientes que no tenían registros en los periodos de observación definidos, esto es el último año (enero- agosto de 2017) y los dos años previos (octubre 2015-marzo 2017, ver sección 8.4).

7.7.2 Etiquetado de casos y controles

El etiquetado de casos en el periodo de selección de la muestra se hizo mediante la identificación de códigos de CIE10 correspondientes a cáncer gástrico (C160 a C169 y D002), y coincidentes con códigos de estudio anatomopatológico (CUPS 8981 y 898101 a 898103). Esto permite identificar pacientes con diagnóstico nuevo de cáncer gástrico reduciendo la probabilidad de tener casos prevalentes (diagnosticados en periodos previos). Los controles se asumieron como los pacientes que no tenían registros con código de cáncer gástrico en el periodo de selección de la muestra.

Adicionalmente, se rastrearon las bases de datos dentro de todo el periodo de información disponible (enero 2011 a agosto 2017), a fin de descartar pacientes que, en fechas previas al periodo de selección, tuvieran registros asociados a los códigos de cáncer gástrico (CIE10 C160 a C169 y D002) o pacientes que tuvieran registros con código de gastrectomía (CUPS 43.7 a 43.9). Esto con el fin de reducir la probabilidad de falsos positivos (diagnóstico de cáncer gástrico previo al periodo de selección) o falsos negativos (antecedente de cáncer gástrico).

7.7.3 Pre-procesamiento de datos (filtros para selección de variables)

El objetivo del modelo de aprendizaje de máquinas es detectar la huella que deja el paciente en el sistema de salud durante el proceso de evolución clínica de la enfermedad y hasta el diagnóstico (enfermedad sintomática), eliminando el sesgo de la selección a

priori de las variables a considerar. Por ello, para el modelo cada código diagnóstico del CIE10 (3 dígitos), cada código CUPS y cada municipio se consideró una variable binaria, pero debido al elevado número de funciones y el costo computacional asociado, fue necesario hacer una reducción de la dimensionalidad. Para este propósito, se aplicaron filtros por diagnóstico y por procedimiento.

En relación con los diagnósticos, se eliminaron códigos concernientes a embarazo, parto y puerperio (códigos CIE10 O00 a O99); códigos relacionados con trauma y otras causas externas (códigos CIE10 de las series S, T, V, W, X, Y), códigos para situaciones especiales (códigos CIE10 U00-U89) y los códigos relacionados con factores que influyen en el estado de salud y el acceso a los servicios, pero que no representan una condición patológica directamente (códigos CIE10 Z00 a Z99).

Una vez filtrados los códigos se corrió un modelo preliminar con regresión logística LASSO, de manera independiente para los dos periodos de análisis (último año y dos años atrás) y para las dos categorías de variables (diagnósticos y procedimientos). Este ejercicio permitió seleccionar variables con mayor capacidad predictiva (coeficiente diferente de cero y OR diferente de 1). Los parámetros para la sintonización del modelo fueron:

Parámetro	Diagnósticos	Procedimientos
Técnica	Train validation split	Train validation split
Razón de entrenamiento*	0,6/0,2	0,6/0,2
Máximo de iteracones	1000	500
Valor de t	1e-10	1e-8
Alfa	0,05	0,05
Lambda de prueba	0,001 hasta 0,05	0,001 hasta 0,05
Pruebas de lambda	cada 0,001	cada 0,001
Lambda resultante un año	0,012	0,010
Lambda resultante dos años	0,016	0,011

La razón de entrenamiento se aplica al 80% de datos seleccionados aleatoriamente para entrenamiento

Con el parámetro de regularización lambda sintonizado, se generaron los modelos para diagnósticos y procedimientos mediante validación cruzada (10 K-folds), con razón de entrenamiento 0,8/0,2, número máximo de iteraciones 1000, valor de t 1e-10 y los valores de lambda resultantes para cada modelo en la fase de sintonización.

A fin de homogenizar los modelos, a partir de los resultados de la regresión se seleccionaron variables usando los siguientes criterios para los dos periodos de análisis:

- Si el coeficiente del diagnóstico (código CIE10) fue diferente de cero y el OR diferente de 1 en al menos uno de los dos periodos, se mantuvo el código para ambos periodos.
- Si el coeficiente para el código CIE10 fue cero y el OR 1 en los dos periodos, se eliminó el código correspondiente exceptuando los códigos de patologías infecciosas y los códigos de la serie R (hallazgos clínicos y de laboratorio anormales). De estos últimos solo se excluyeron los códigos para síntomas generales (síntomas de la infancia, dedos deformes, síntomas no especificados), y causas de morbilidad desconocidas (CIE10 R68-R69).

Finalmente, los códigos seleccionados se agruparon a fin de mejorar su capacidad predictiva dada la colinealidad de los diagnósticos y la elevada probabilidad de asignación de códigos indiferenciados al interior de los grupos establecidos (Tabla 14).

Los procedimientos se agruparon dejando solamente los primeros 4 dígitos del código CUPS y se utilizó el mismo filtro; es decir, si el coeficiente para el código CUPS fue diferente de cero con OR diferente de 1, se mantuvo exceptuando los siguientes códigos:

- Consulta (CUPS 8900, 8903)
- Actividad educativa (CUPS 990, 9901, 9902)
- Pruebas rutinarias (CUPS 9022)
- Procedimientos especiales (CUPS 9068)
- Sustitución de férula (CUPS 9711 y 9712). Relación con trauma

Tabla 14. Agrupación de códigos diagnósticos (CIE10) para la construcción del modelo basado en aprendizaje de máquinas

Infecciones	Parasitosis	Anemia	Enfermedades endocrinas, nutricionales y metabólicas	Trastornos mentales	Enfermedades del ojo y sus anexos	Enfermedades del sistema circulatorio
A00-A19	B00-B09	D50-D53	E40-E43	F00-F09	H65-H70	I10-I15
A15-A19	B15-B19		E50-E56		H90-H91	I20-I25
A30-A49	B25-B34					I26-I28
A80-A89	B35-B49					I44-I45
	B50-B64					I80-I82
	B65-B83					

7.7.4 Procesamiento: desarrollo de los modelos de aprendizaje de máquinas

Se utilizaron tres modelos de aprendizaje de máquinas: LASSO, bosques aleatorios y potenciación del gradiente. Los modelos se representan en la Figura 4.

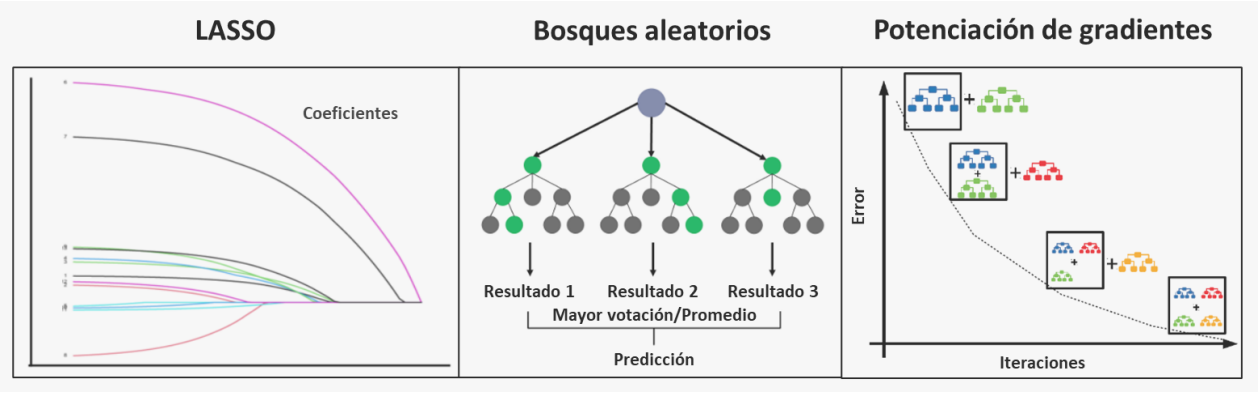


Figura 4. Representación esquemática de los modelos de aprendizaje de máquinas utilizados en el estudio

1. Como se describió en la sección 8.6.3, LASSO se utilizó para la selección de variables debido a que combina un modelo de regresión con un procedimiento de

contracción imponiendo una restricción sobre los coeficientes de regresión (67). La reducción de los coeficientes hace que se reduzca el problema de colinealidad lo cual es una ventaja importante cuando se tiene gran cantidad de parámetros como en este estudio. A diferencia de otros métodos, LASSO penaliza también la pendiente del modelo (lambda), con esta medida el coeficiente de los parámetros menos importantes tiende a cero, lo cual permite una mejor selección de las variables que no tienen capacidad predictiva y hace el modelo más fácil de interpretar.

Una vez seleccionadas las variables a incluir (Ver sección 8.6.3), se corrió de nuevo el modelo utilizando el 80% de los datos para entrenamiento y el 20% para prueba del modelo entrenado. Se hizo de nuevo la definición del parámetro lambda con la base de datos completa (sintonización sin separación de diagnósticos y procedimientos) utilizando el 60% de los datos y se validó con el 20% restante del conjunto de entrenamiento, siguiendo los mismos parámetros utilizados para la selección de variables. El lambda resultante fue 0,009 para los dos periodos de análisis. Se realizaron 5 validaciones cruzadas (K-folds) con 10 subconjuntos de datos. El ajuste de pesos (balanceo) se hizo con la función *"balanced"* en lugar de *"none",* el máximo de iteraciones fue 100 dada la sintonización previa y valor de t utilizado (1e-4), esto a fin de reducir la exigencia de recursos computacionales; en este caso el número máximo de iteraciones no se alcanzó. La razón de entrenamiento fue 0,6/0,2 para el 80% de los datos en el set de entrenamiento.

La regresión logística LASSO se utilizó para los dos periodos de análisis (último año y dos años atrás). Los resultados obtenidos con la regresión logística LASSO mostraron mayor capacidad predictiva con el modelo basado en datos del periodo de análisis más cercano al desenlace (último año), por ello, para efectos de comparación se decidió utilizar los otros modelos de aprendizaje de máquinas únicamente para el periodo de dos años atrás.

2. La predicción por bosques aleatorios selecciona inicialmente un subconjunto aleatorio de datos y genera numerosos árboles de decisión (68). Luego resume los resultados de los diferentes árboles de decisión tomando como predicción de la variable de desenlace el resultado reportado por la mayoría de los árboles construidos. Esto reduce el error frente a la decisión tomada por un solo árbol. El algoritmo utiliza valores ponderados al considerar el error de cada resultado en los árboles de decisión; así, un árbol con un error alto recibe un peso bajo y un árbol con un error bajo recibe un peso alto.

   En la construcción del modelo se utilizó la base de datos con las variables seleccionados mediante LASSO y correspondiente al periodo de análisis de dos años atrás. Se utilizó el 70% de la muestra para entrenamiento y el 30% restante para prueba del modelo. Se realizaron 5 validaciones cruzadas (K-folds) cada una con 10 subconjuntos de datos. Se utilizaron los parámetros de la máquina por defecto, pero se definieron los siguientes:

   - Número de árboles 2000

- Número mínimo de submuestras 1
- Número mínimo de características a nivel de nodo “sqrt”
- Profundidad máxima 5
- Ganancia mínima de información cross validation
- Criterio para ganancia de información Gini

El número de árboles se definió en 2000 a fin de tener un número mayor que el número de variables, pero con un costo computacional factible. Por la misma razón el número mínimo de submuestras se definió en 1 y el número mínimo de características (variables) en el nodo se estableció como “sqrt” (raíz cuadrada del total de variables) en lugar de usar la función “auto”, y se utilizó la función “Gini” (lineal) en lugar de la función “entropy” (logarítmica). La profundidad máxima se estableció en 5 con base en la revisión de literatura en donde se sugiere que este número permite evitar el sub- o sobre-ajuste.

3. La potenciación del gradiente se basa igualmente en árboles de decisión, pero en lugar de hacer predicciones independientes en cada árbol, ajusta cada nuevo árbol en un proceso de aprendizaje basado en los datos históricos (68, 69). De esta forma, si un árbol no explica completamente las variaciones en la variable de desenlace, se calcula el error y en la próxima iteración se crea un nuevo árbol de decisión que explica las variaciones con una variancia residual. Este proceso se ejecuta varias veces, de manera que los errores más pequeños y los errores residuales más pequeños se utilizan como entradas en los árboles adicionales.

   Igual que en los bosques aleatorios, se utilizó la base de datos con las variables seleccionadas mediante LASSO y correspondiente al periodo de análisis de dos años atrás. En la construcción del modelo se utilizó el 70% de la muestra para entrenamiento y el 30% restante para prueba. Se realizaron 5 validaciones cruzadas (K-folds) cada una con 10 subconjuntos de datos. Se utilizaron los parámetros de la máquina por defecto:

   - Número máximo de iteraciones 20
   - Número mínimo de submuestras 1
   - Número mínimo de características a nivel de nodo “all”
   - Profundidad máxima 5
   - Ganancia mínima de información cross validation
   - Criterio para ganancia de información “variance”
   - Tasa de aprendizaje 0,1
   - Umbral de validación (tol) 0,01
   - Función de pérdida “logistic”

   El bajo número de iteraciones es posible dada la elección secuencial de árboles con base en la varianza (“variance”). En términos de recursos computacionales ello compensa la inclusión de la totalidad de variables a nivel de nodo (“all” para submuestra mínima 1) y la tasa de aprendizaje lenta, con la cual se espera un modelo más robusto. El modelo ajustó por la función de pérdida (“logistic” para desenlace binario) y no llegó al total de iteraciones.

7.7.5 Evaluación de los modelos predictivos

La exactitud de los tres modelos construidos se evaluó a partir de la determinación del área bajo la curva. Además, con base en la matriz de confusión se calculó

- Verdaderos positivos (VP): número de casos con diagnóstico de cáncer gástrico clasificados como cáncer gástrico por el modelo
- Verdaderos negativos (VN): número de controles clasificados como controles por el modelo (sin diagnóstico de cáncer gástrico)
- Falsos positivos (FP): número de controles clasificados como casos de cáncer gástrico por el modelo
- Falsos negativos (FN): número de casos de cáncer gástrico clasificados como controles por el modelo
- Exactitud: porcentaje de tumores cuya malignidad se predijo correctamente.
- Sensibilidad: VP / (VP + FN)
- Especificidad: VN / (VN+FP)
- Valor predictivo positivo (VPP): VP / (VP+FP)
- Valor predictivo negativo (VPN): VN / (VN+FN)

Adicionalmente se calculó el número de individuos necesario a seguir en función de la probabilidad de diagnóstico de cáncer; es decir, cuantos sujetos deberían entrar en seguimiento clínico para diagnosticar un caso de cáncer en un periodo de un año (modelo último año) o dos años (modelo dos años atrás).

8. Resultados

8.1 Modelo multivariante

Se identificaron 14399 individuos con anemia y 44961 individuos sin anemia en 2011. La edad fue igual entre los dos grupos, pero la población con anemia fueron más frecuentemente mujeres y residentes en ciudades capitales (Tabla 15).

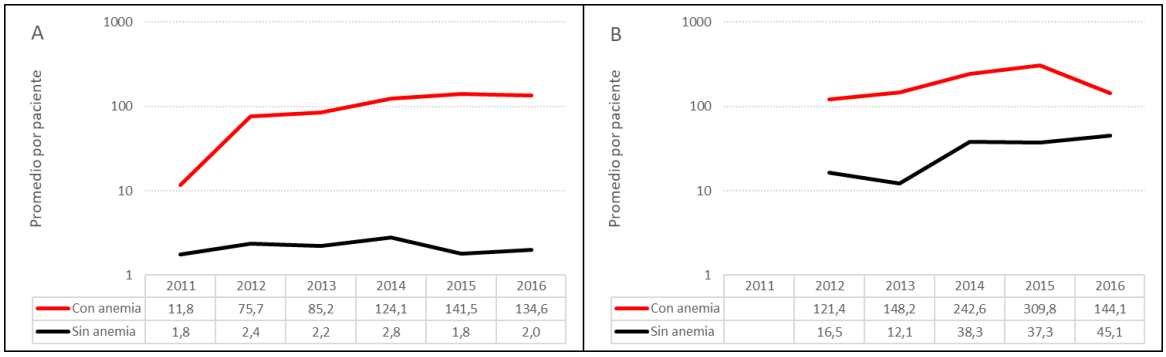
Tabla 15. Características de la población de estudio para construcción del modelo multivariante

Variable	Con anemia n=14399	Sin anemia n=44961
Edad, media (rango)	59 (51-76)	59 (52-59)
Sexo, No. (%)		
Femenino	11.781 (88,1%)	27.008 (60,1%)
Masculino	2.618 (18,2%)	17.953 (39,9%)
Municipio, No. (%)		
Capital	11.058 (76,8%)	31.364 (69,8%)
Resto	3.341 (23,2%)	13597 (30,2%)
Comorbilidad, mediana (rango IC)	7 (6-8)	6 (6-7)

La comorbilidad tuvo puntuación entre 6 y 20, pero no se observaron mayores diferencias entre los pacientes con y sin anemia. El promedio de atenciones fue 1,8; 23,0 y 40,3 para

los pacientes sin anemia, con anemia y con cáncer gástrico, respectivamente. Las diferencias entre sujetos con anemia y sin anemia se observaron inclusive en los pacientes diagnosticados con cáncer gástrico; sin embargo, en esta población la diferencia entre los dos grupos fue menor (Gráfico 3). Se observó una tendencia al incremento en el promedio de atenciones en la población con anemia y los pacientes con cáncer gástrico, la población sin anemia mantuvo una tendencia estable.

Gráfico 3. Tendencia del uso de servicios (atenciones) en la población de estudio durante el periodo de observación



A. Promedio sobre el total de pacientes en cada grupo B. Promedio sobre los pacientes diagnosticados con cáncer gástrico en cada grupo. Los promedios se establecen sobre el número de pacientes que tuvieron atenciones en cada año. La escala está expresada en valores logarítmicos.

El riesgo relativo (no ajustado) observado para el diagnóstico de cáncer gástrico en pacientes con diagnóstico de anemia comparado con los pacientes sin anemia fue 1,9 (IC95% 1,5-2,4), con fracción atribuible en la población expuesta 47,1 y fracción atribuible poblacional 17,8 (Tabla 16). No se observaron diferencias en la evaluación de la asociación cuando se usó únicamente el diagnóstico de cáncer gástrico versus la inclusión de muertes por cáncer gástrico en la variable de desenlace.

Tabla 16. Medidas de asociación (valores no ajustados) del antecedente de anemia con el diagnóstico de cáncer gástrico

Estimador	Desenlace: cáncer gástrico	
	Diagnóstico valor (IC95%)	Diagnóstico/Muerte valor (IC95%)
Riesgo relativo	1,9 (1,5-2,4)	1,9 (1,5-2,4)
Riesgo atribuible	0,4 (0,2-0,6)	0,4 (0,3-0,6)
Riesgo atribuible poblacional	0,1 (0,0-0,2)	0,1 (0,0-0,2)
Fracción atribuible	47,1 (33,9-57,8)	47,4 (34,6-57,7)
Fracción atribuible poblacional	17,8 (10,6-24,4)	17,9 (10,9-24,4)

En la columna titulada "Diagnóstico" se asumió la variable de desenlace considerando únicamente los códigos CIE10 de cáncer gástrico (C160-C169 y D002) asociados a las atenciones registradas en la base de datos. En la columna titulada "Diagnóstico/Muerte" se incluyeron como parte del desenlace las muertes por cáncer gástrico que no tenían previamente este diagnóstico.

Los valores ajustados se muestran para el desenlace Diagnóstico/Muerte por cáncer gástrico. Todas las variables en el modelo mostraron asociación excepto la edad (Tabla

17). El OR mas alto se observó para el sexo masculino (OR 1,68), seguido de el antecedente de anemia (OR 1,53).

Tabla 17. Medidas de asociación ajustadas para las variables incluidas en el modelo

Variable	OR	IC95%	
		LI	LS
Anemia	1,53	1,51	1,54
Comorbilidad	1,35	1,35	1,35
Sexo (Masculino)	1,68	1,67	1,69
Edad	1,03	1,03	1,03

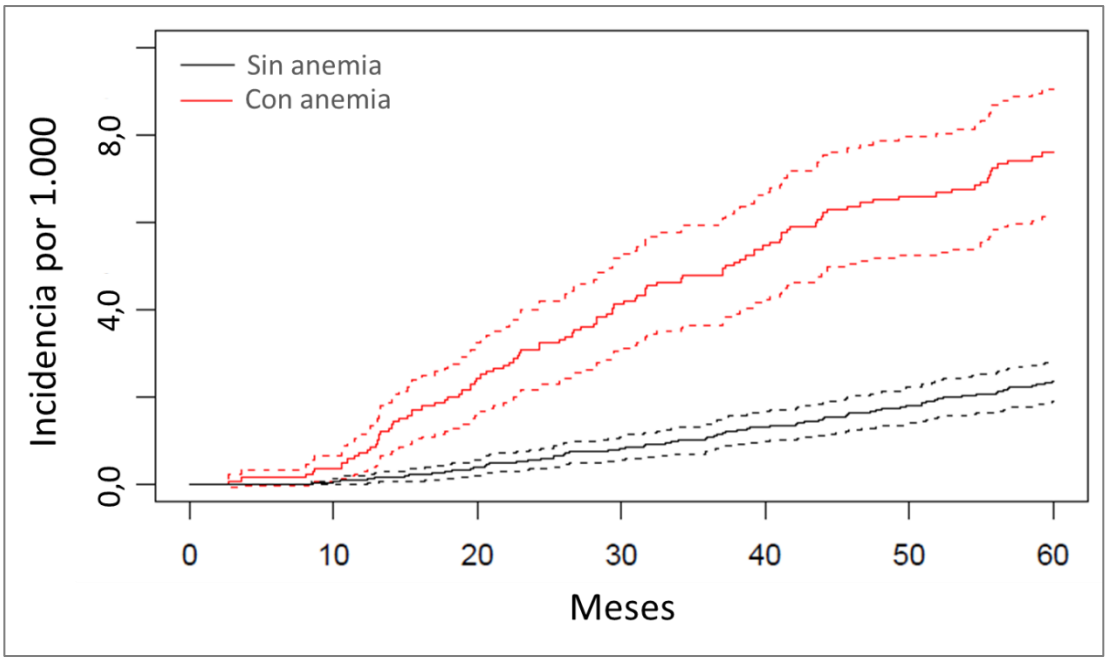
Residuos de desviación Min -0,94 1Q -0,11 Mediana -0,09 3Q -0,07 Max 3,54. AIC 3924,5

La evaluación de la capacidad predictiva del modelo mostró los siguientes parámetros:

- Sensibilidad 71,3% (IC95% 69,4-73,6)
- Especificidad 72,4% (IC95% 70,0-74,3)
- Exactitud (*accuracy*) 72,4% (IC95% 70,2-74,3)
- Área bajo la curva 0,78 (IC95% 0,75-0,80)

En el modelo de tiempo al evento, la mediana de tiempo al diagnóstico para los sujetos con anemia fue 28,2 meses (RIC 17,2-40,9), para los sujetos sin anemia 36,7 meses (RIC 23,8-48,0) y para toda la población de estudio 31,6 meses (RIC 19,9-44,1). La mortalidad general observada fue 190,0 por 100.000. En total 1311 sujetos que iniciaron sin anemia fueron diagnosticados con anemia después de 2011.

Gráfico 4. Incidencia acumulada de cáncer gástrico en sujetos con y sin anemia. Cálculo según modelo de riesgos competitivos



La incidencia de cáncer gástrico acumulada a 60 meses en los sujetos con anemia fue 7,6 por 1.000 (IC95% 6,1-9,1) y en los sujetos sin anemia 2,4 por 1.000 (IC95% 1,9-2,8)

(Gráfico 4). En el modelo de riesgos competitivos la anemia mostró la mayor fuerza de asociación seguida del sexo masculino (Tabla 18).

Tabla 18. Razón de riesgos de cáncer gástrico para las variables en el modelo de riesgos competitivos según Fine y Gray

Variable	HR	IC95%	
		LI	LS
Anemia	2,95	2,18	3,99
Comorbilidad	1,27	1,21	1,32
Sexo (Masculino)	2,10	1,58	2,80
Edad	1,03	1,02	1,04

8.2 Modelo de aprendizaje de máquinas

Luego de filtros iniciales (sin dato nacimiento, duplicado ID, edad 30-89), se obtuvieron 5.768.231 IDs único, de los cuales 1.507 corresponden a casos de cáncer (Figura 5).

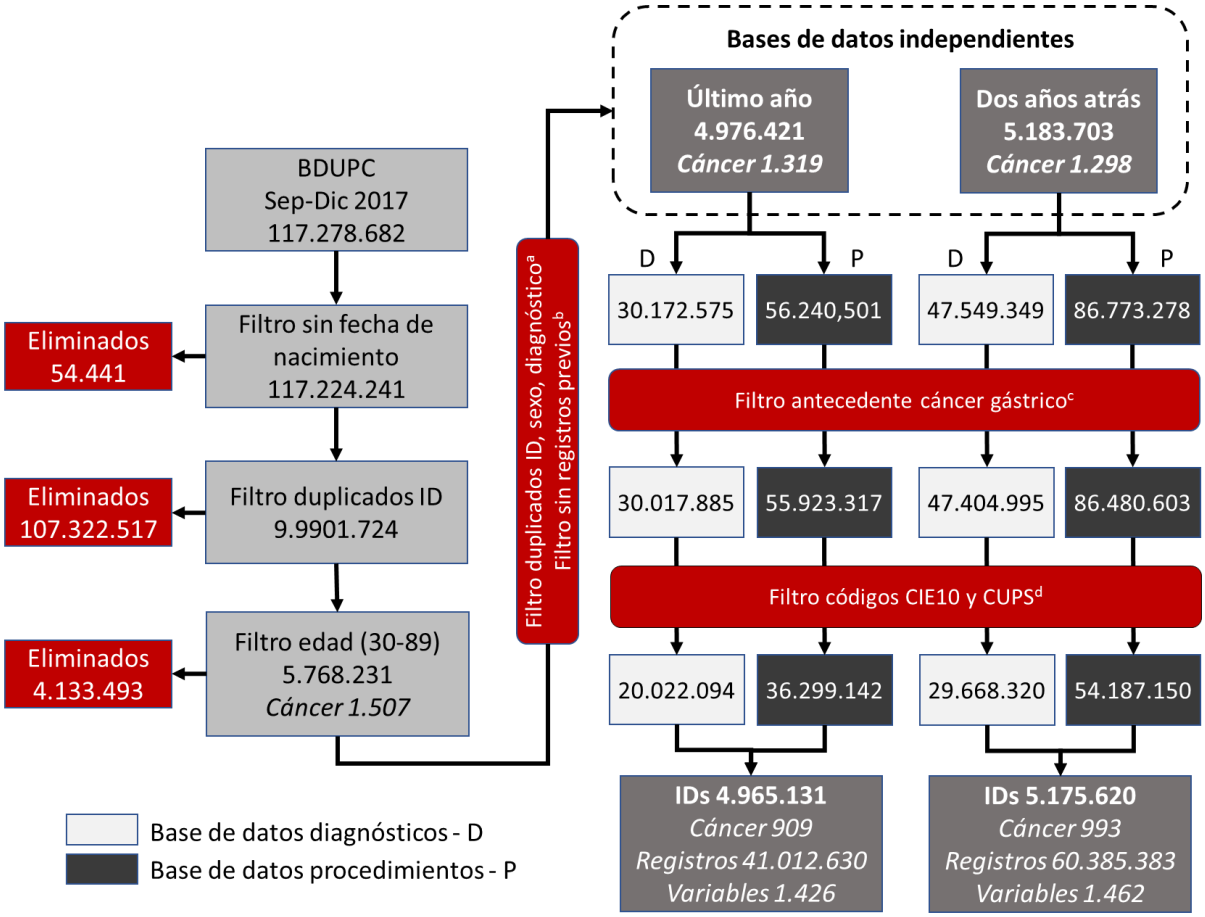


Figura 5. Flujograma del proceso de selección de la muestra en aprendizaje de máquinas

a) Filtro de registros que aportan la misma información por ser duplicados de las tres variables. b) Filtro de IDs que no tenían registros de atención antes de septiembre de 2017 (no aportan datos de seguimiento). c) Filtro de IDs que tenían códigos CIE10 o CUPS asociados a cáncer gástrico entre enero de 2011 y agosto

de 2017 (mejora de etiquetado). d) Filtro de registros que no tenían variables con capacidad predictiva (coeficiente cero); incluye el agrupamiento de códigos CIE10 y CUPS.

Tras los filtros por antecedentes de cáncer gástrico en periodos anteriores (Código CIE10 o CUPS asociados a cáncer gástrico entre enero 2011 y agosto 2017) y el proceso de selección de variables predictoras, se obtuvieron 4.965.931 y 5.175.620 registros con ID único en las bases de datos para seguimiento el último año y los dos años previos, respectivamente; dentro de los cuales hubo 1.426 y 1.462 casos de cáncer gástrico (Figura 5).

En total hubo 909 y 993 casos de cáncer gástrico en la base de datos de seguimiento el último año y los últimos dos años respectivamente, frente a 4.964.222 y 5.174.627 pacientes sin cáncer en cada base de datos respectivamente. La mayoría de la población fueron mujeres tanto en los pacientes con cáncer como en los controles sin cáncer, sin variación en los dos periodos de análisis (Tabla 19). La edad promedio fue 58 años en los pacientes con cáncer gástrico y 53 años en los controles sin cáncer.

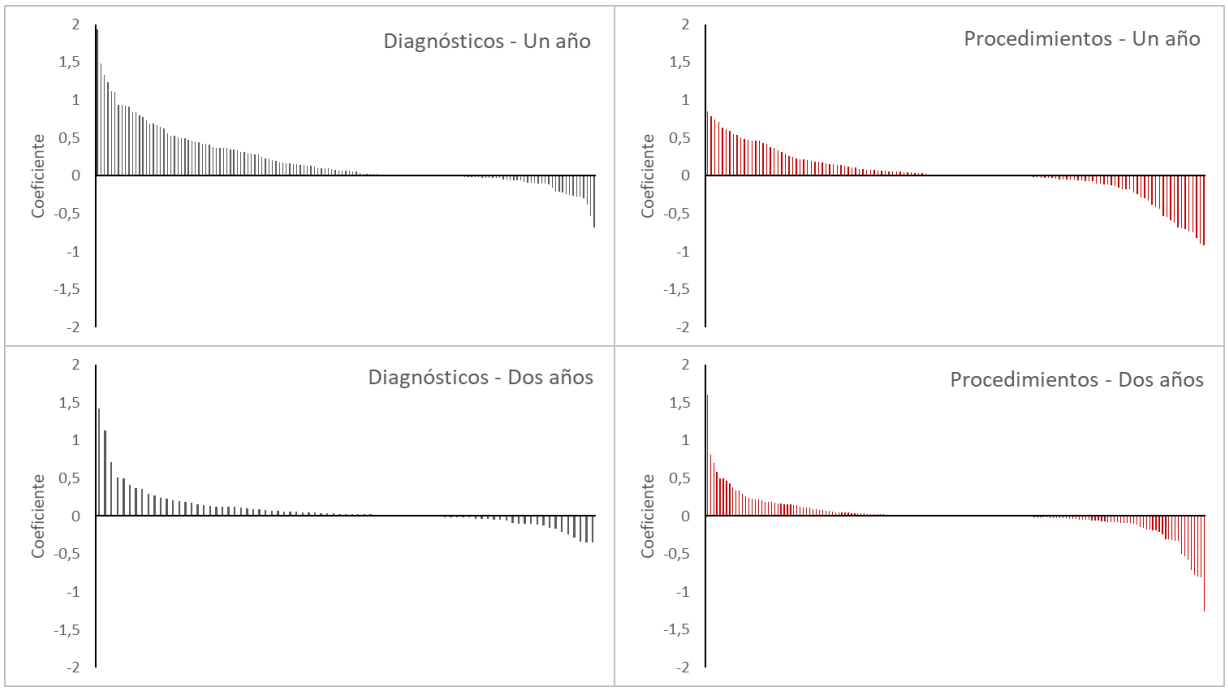
Tabla 19. Características de la población de estudio en los modelos de aprendizaje de máquinas según periodo de análisis

Variable	Cáncer gástrico	Controles sin cáncer
Un año antes del diagnóstico		
No. individuos (IDs)	909	4.964.222
No. Registros (total)	11.229	41.001.401
Tasa de atenciones (registros/individuos)	12,4	8,3
Edad (%)		
30-39	93 (10,2%)	1.186.735 (23,9%)
40-44	62 (6,8%)	484.279 (9,8%)
45-90	754 (82,9%)	3.293.208 (66,3%)
Promedio de Edad	57,6	53,0
Sexo (%)		
Hombre	374 (41,1%)	1.863.012 (37,5%)
Mujer	535 (58,9%)	3.101.210 (62,5%)
No. Aseguradores (# EPS)	9	11
No. Diagnósticos (# códigos CIE10)	157	167
No. Procedimientos (# códigos CUPS)	123	135
Dos años antes del diagnóstico		
No. individuos (IDs)	993	5.174.627
No. Registros (total)	15.426	60.369.957
Tasa de atenciones (registros/individuos)	15,5	11,7
Edad (%)		
30-39	101 (10,2%)	1.277.517 (24,7%)
40-49	71 (7,2%)	516.147 (10,0%)
50-90	821 (82,7%)	3.380.963 (65,3%)
Promedio de Edad	57,8	52,7
Sexo (%)		
Hombre	425 (42,8%)	1.969.563 (38,1%)
Mujer	568 (57,2%)	3.205.064 (61,9%)
No. Aseguradores (# EPS)	13	15
No. Diagnósticos (# códigos CIE10)	150	173
No. Procedimientos (# códigos CUPS)	144	154

El uso de servicios (tasa de atenciones) fue mayor en pacientes con cáncer gástrico en ambos periodos, pero la diferencia con pacientes sin cáncer fue mayor en el año previo al diagnóstico. No obstante, el número de diagnósticos y procedimientos fue menor en pacientes con cáncer gástrico.

Los coeficientes para diagnósticos y procedimientos incluidos como variables predictivas fueron mayores en el análisis del último año, indicando mayor capacidad predictiva en la medida en que el desenlace es más cercano en el tiempo (Gráfico 5). A su vez, los diagnósticos mostraron mayor capacidad predictiva en ambos periodos y los coeficientes de los procedimientos mostraron valores más balanceados entre resultados positivos y negativos que los diagnósticos, indistintamente del periodo de análisis.

Gráfico 5. Distribución de los coeficientes para las variables incluidas en el modelo (diagnósticos y procedimientos)



El área bajo la curva ROC (AUROC) para los modelos LASSO, bosques aleatorios y potenciación de gradientes no tuvo diferencias significativas en el análisis de los dos años previos al diagnóstico de cáncer (Tabla 20); sin embargo, medidos en el punto de corte 0,5 (probabilidad de predicción del desenlace), la exactitud fue, desde el punto de vista estadístico, significativamente menor en el modelo de regresión logística LASSO que en los otros dos modelos. A pesar de ello, la sensibilidad en el punto de corte con mayor exactitud para LASSO es mayor que la sensibilidad en el punto de corte con mayor exactitud en bosques aleatorios y potenciación de gradiente (65,4%; 27,8% y 57,9%, respectivamente) y la especificidad solamente fue significativamente mayor en el modelo de bosques aleatorios, pero a costa de una reducción significativa en la sensibilidad.

La capacidad predictiva del modelo de regresión logística LASSO en el último año previo al diagnóstico fue significativamente mayor que los modelos de predicción para los dos años previos al diagnóstico, tanto en el área bajo la curva ROC, como en exactitud, sensibilidad y especificidad.

Tabla 20. Desempeño de los modelos de aprendizaje de máquinas en el estudio

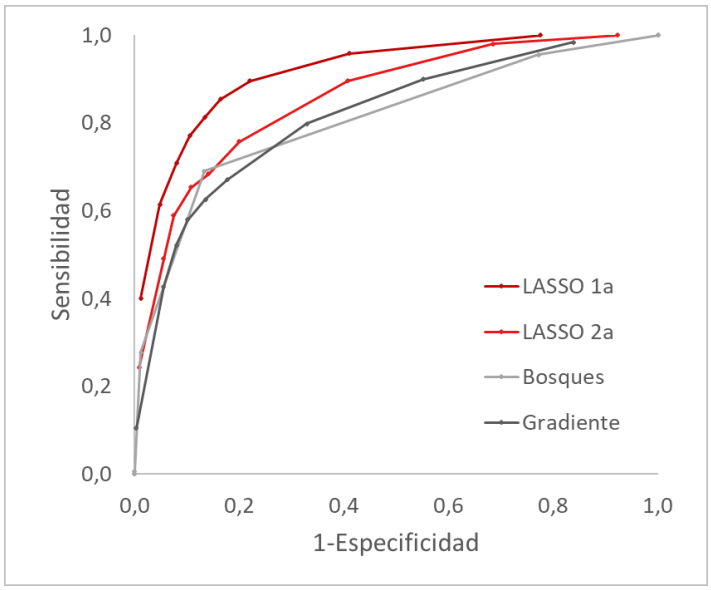
Modelo	AUROC	Exactitud (IC95%)	Sensibilidad (IC95%)	Especificidad (IC95%)	VPP (IC95%)	VPN (IC95%)
LASSO 1 año						
Corte 0,5	0,90 interno	86,6 (86,5-86,6)	81,3 (75,0-86,5)	86,6 (86,5-86,6)	0,12 (0,11-0,13)	100,0 (99,0-100,0)
Corte 0,7	0,92 externo	92,0 (91,9-92,1)	70,8 (63,9-77,2)	92,0 (91,9-92,1)	0,17 (0,16-0,19)	100,0 (99,0-100,0)
LASSO 2 años						
Corte 0,5	0,84 interno	85,9 (85,7-85,9)	68,3 (61,4-74,7)	85,9 (85,7-85,9)	0,09 (0,09-0,10)	100,0 (99,0-100,0)
Corte 0,6	0,87 externo	89,2 (89,2-89,3)	65,4 (58,4-71,9)	89,2 (89,2-89,3)	0,12 (0,11-0,13)	100,0 (99,0-100,0)
Bosques aleatorios						
Corte 0,5	0,91 interno	86,7 (86,6-86,7)	69,0 (63,6-74,1)	86,7 (86,6-86,7)	0,10 (0,10-0,11)	100,0 (99,0-100,0)
Corte 0,6	0,85 externo	98,7 (98,7-98,8)	27,8 (22,9-33,1)	98,7 (98,7-98,8)	0,45 (0,37-0,53)	100,0 (99,0-100,0)
Potenciación de gradiente						
Corte 0,5	0,92 interno	86,4 (86,4-86,5)	62,6 (56,9-68,2)	86,4 (86,4-86,5)	0,09 (0,08-0,10)	100,0 (99,0-100,0)
Corte 0,6	0,82 externo	89,8 (89,8-89,9)	57,9 (52,1-63,6)	89,8 (89,8-89,9)	0,11 (0,10-0,12)	100,0 (99,0-100,0)

Cada modelo se corrió 5 veces en el conjunto de datos de prueba (Ver anexos); para cada moldeo se presentan los datos de la prueba con el mayor valor del área bajo la curva ROC. Se presenta el punto de corte con probabilidad 0,5 en la predicción de la enfermedad para todos los modelos y el punto de corte en el que cada modelo mostró la mayor exactitud en la discriminación de pacientes con cáncer gástrico y pacientes sin cáncer gástrico. El valor del área bajo la curva ROC (AUROC por su sigla en inglés) se presenta frente al conjunto de datos de entrenamiento (interno) y el conjunto de datos de prueba (externo). VPP: Valor predictivo positivo. VPN: Valor predictivo negativo. El reporte para todos los puntos de corte se encuentra en los anexos.

En todos los casos, el desempeño se mejoró con el incremento del punto de corte (mayor probabilidad de predicción del desenlace), como se deduce de los valores para el área bajo la curva ROC y la exactitud; sin embargo, esta mejoría se dio esencialmente por un incremento de la especificidad con detrimento de la sensibilidad. En consonancia con ello, la exactitud y la especificidad reportan valores mínimos (centésimas), sugiriendo que los modelos privilegian la reducción de falsos positivos sobre la predicción del diagnóstico de cáncer. No obstante, debido a la baja prevalencia de la enfermedad (aproximadamente 0,02%), los valores predictivos positivos en todos los modelos fueron menores de 1%. Los valores del área bajo la curva ROC no mostraron diferencias relevantes entre el ejercicio de validación interno (entrenamiento) y la validación externa.

El desempeño de los modelos se evaluó también mediante la construcción de las curvas ROC (Receiving Operating Characteristic), en donde igualmente se observa la mayor capacidad discriminativa del modelo de regresión logística LASSO en el último año y la similitud de los modelos que analizaron dos años atrás (Gráfico 6).

Gráfico 6. Curva ROC de los modelos de aprendizaje de máquinas en el estudio



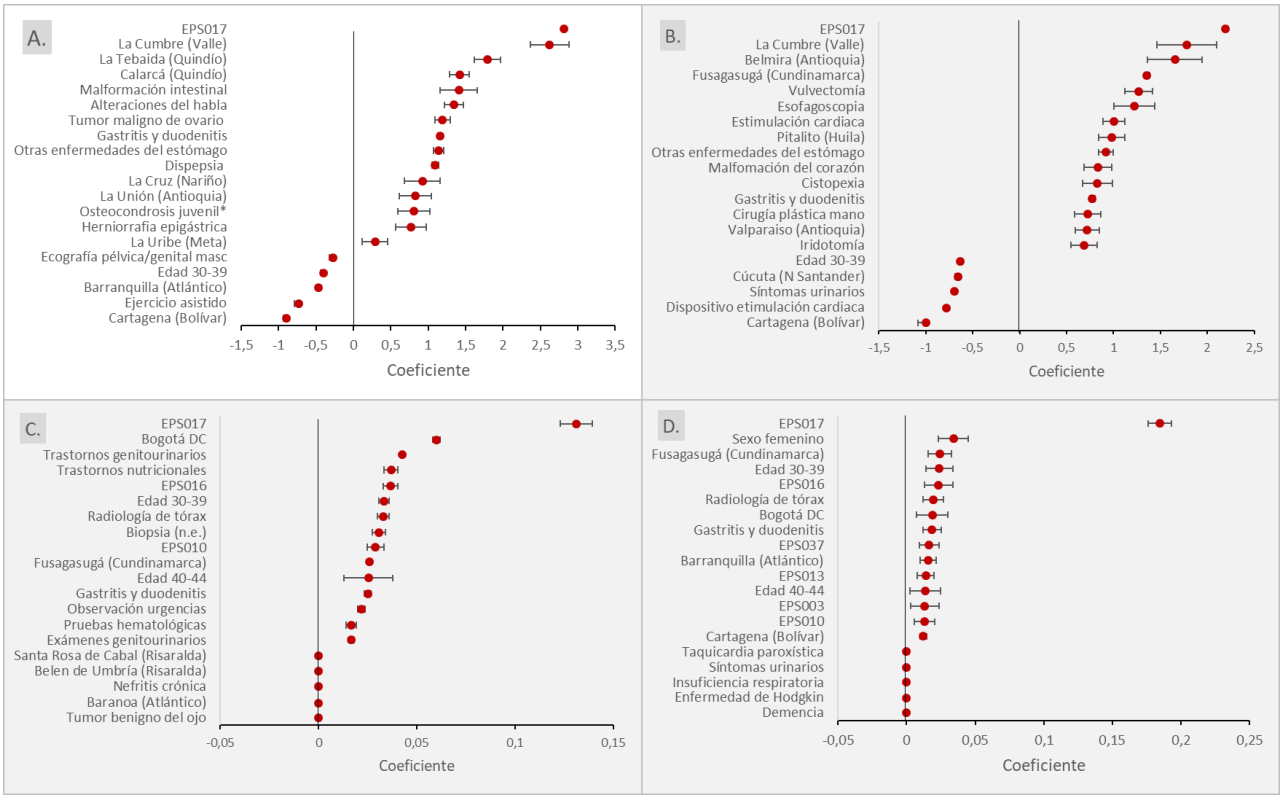
LASSO 1a corresponde a regresión logística LASSO analizando información del último año previo al diagnóstico y LASSO 2a utiliza la información de dos años previos al diagnóstico. Bosques corresponden a bosques aleatorios y Gradiente a potenciación del gradiente, los que también analizan dos años previos al diagnóstico.

El análisis de la capacidad discriminativa de las variables incluidas en el modelo es complejo debido al elevado número utilizado (1.426 y 1.462 para los análisis a un año y dos años atrás, respectivamente – Figura 5). Una revisión de las variables con los mayores y los menores valores del coeficiente muestra resultados disímiles entre los modelos (Gráfico 7).

Solo los modelos de regresión logística LASSO reportaron variables con coeficiente negativo. En todos los modelos la aseguradora identificada como EPS017 mostró el coeficiente más alto. Todos los modelos reportan municipios ubicados en la zona de cordillera dentro de las variables con coeficientes más altos y municipios ubicados en las

planicies del norte del país dentro de las variables con coeficientes más bajos. La edad menor de 40 años se reporta con valor negativo en los modelos de regresión logística LASSO (reducción de riesgo); sin embargo, se reporta con valor positivo dentro de los coeficientes más altos en los modelos de bosques aleatorios y potenciación de gradientes.

Gráfico 7. Contribución de las variables con los mayores y menores coeficientes a la capacidad predictiva de los modelos de aprendizaje de máquinas



A. LASSO último año. B. LASSO dos años atrás. C. Bosques aleatorios. C. Potenciación de gradiente. B, C y D analizan dos años atrás. Se presentan las 15 variables con coeficiente mas alto y las 5 variables con coeficiente mas bajo para cada uno de los modelos. El total de variables es 1.426 y 1.462 para los análisis del último año y dos años atrás, respectivamente

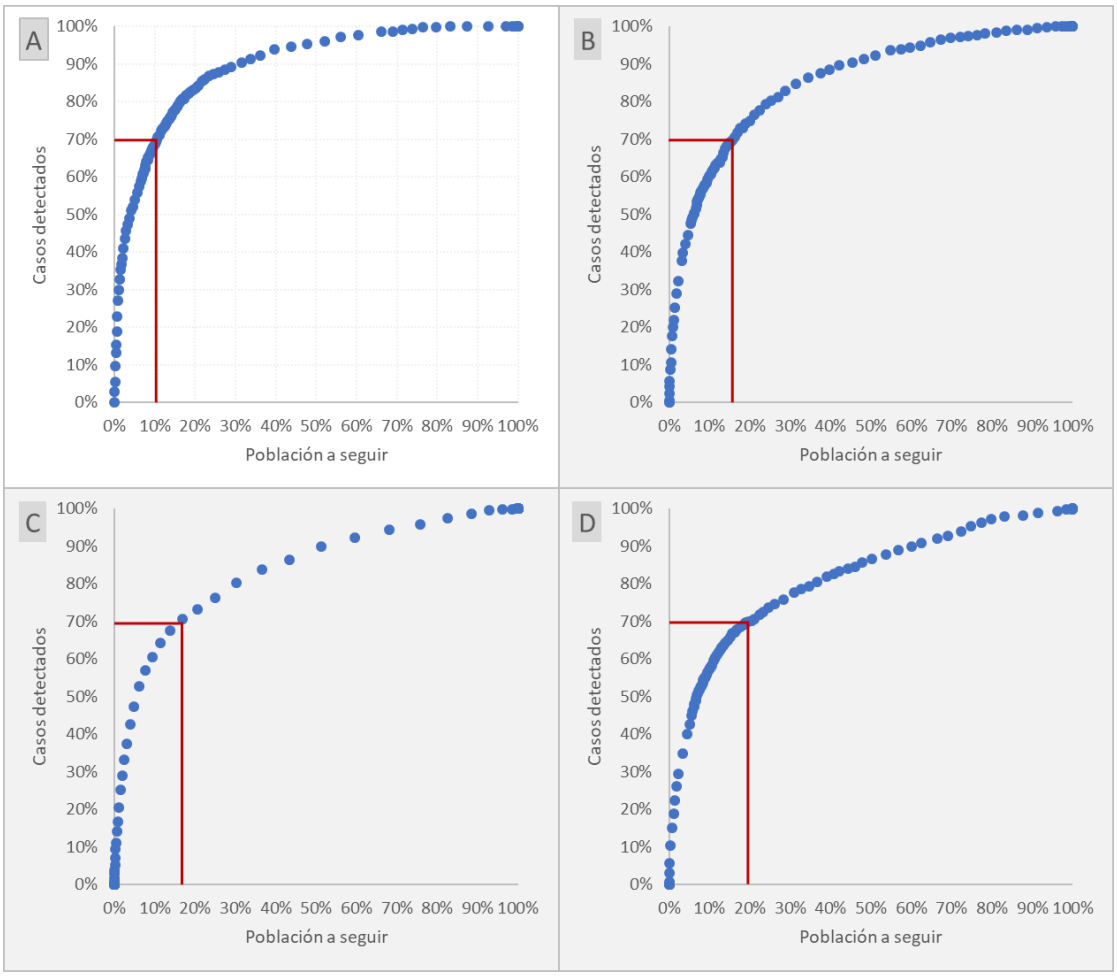
El modelo de regresión logística LASSO del último año reporta más diagnósticos y más municipios dentro de las variables con coeficiente más alto y el modelo de potenciación de gradientes reporta solo un diagnóstico dentro de las 15 variables con mayor coeficiente (Gastritis y duodenitis). Los síntomas y enfermedades gastrointestinales prevalecen en todos los modelos y las anemias carenciales solo se reportan dentro de los principales diagnósticos en el modelo de regresión logística LASSO para el último año (ver anexos), pero no aparece dentro de las principales variables en la mirada general del modelo.

En el modelo de regresión logística LASSO dos años atrás del diagnóstico predominan los procedimientos y los municipios en las variables con coeficiente más alto y en el modelo de potenciación de gradientes los procedimientos y las aseguradoras.

Información más detallada sobre al aporte de las variables de diagnóstico, de procedimiento, de municipio y aseguradoras se presenta en los anexos.

La estimación de la proporción de individuos que requieren seguimiento para detectar al menos el 70% de los casos de cáncer gástrico es 16,1%; 16,5%; y 20,4% en los modelos con análisis dos años atrás del diagnóstico (LASSO, bosques aleatorios y potenciación del gradiente, respectivamente - Gráfico 8); esto equivale a un número necesario a tratar de 1.170, 1.230, y 1.498 personas para cada modelo. El modelo de regresión logística LASSO a un año reporta que se debe seguir el 10,6% de la población para detectar al menos el 70% de los casos de cáncer gástrico, lo que equivale a un número necesario a tratar de 790 individuos (información detallada en los anexos).

Gráfico 8. Relación entre población a seguir y porcentaje de casos de cáncer gástrico detectados según los modelos de aprendizaje de máquinas en el estudio



A. LASSO último año. B. LASSO dos años atrás. C. Bosques aleatorios. C. Potenciación de gradiente. B, C y D analizan dos años atrás.

9. Discusión

América Latina es la segunda región con mayor incidencia y mortalidad por cáncer gástrico a nivel global (22). En consonancia con ello, la actividad investigativa regional ha contribuido de manera significativa al entendimiento de la epidemiología, la biología y

la historia natural de la enfermedad (95). Sin embargo, la investigación sobre detección temprana es escasa en la región, y en la revisión realizada, no se encontraron trabajos previos basados en el uso de inteligencia artificial en este campo a pesar del auge de esta alternativa tecnológica.

Colombia es un país de riesgo medio para cáncer gástrico en el escenario global (incidencia 12,9 por 100.000)(22); no obstante, al interior del país las tasas oscilan entre 2,5 por 100.000 para las mujeres en Sucre y 29,7 para los hombres en Huila (89). Igual que para otros países de la región, tener un programa de tamización de base poblacional o selectivo de población de alto riesgo es poco factible dada la incertidumbre de la evidencia y los requerimientos institucionales expresados en el marco conceptual. Como alternativa, la vigilancia endoscópica de lesiones precancerosas se recomienda ampliamente; sin embargo, tampoco existe evidencia de su eficacia (21), y su efecto en la reducción de la mortalidad es incierto, toda vez que la detección de dichas lesiones depende de los síntomas gastrointestinales que indican la remisión a endoscopia. Como se anotó, la frecuencia de sintomatología gastrointestinal en etapas tempranas de la enfermedad no es elevada; de hecho, los síntomas considerados de alarma no se asocian con diagnóstico temprano del cáncer gástrico y la mayoría de casos vienen de pacientes con síntomas que no son considerados de alarma (51, 96).

En este estudio los síntomas gastrointestinales no mostraron un elevado aporte a la capacidad predictiva de los modelos. Es posible que la relación de estos síntomas con múltiples patologías induzca un bajo valor predictivo para el cáncer gástrico (53); sin embargo, debe tenerse en cuenta que en el estudio se utilizó una codificación que privilegia el diagnóstico de enfermedad sobre la descripción de síntomas (CIE10). En este orden de ideas, la gastritis, duodenitis y dispepsia estuvieron siempre dentro de las variables con mayor aporte a la capacidad predictiva en los modelos de aprendizaje de máquinas (Gráfico 7), y adicionalmente, el uso de algunos medicamentos y procedimientos potencialmente relacionados con sintomatología de enfermedad del tracto gastrointestinal superior mostró algún grado de asociación en el modelo multivariante (Anexo 1).

De otra parte, en la selección de variables para el modelo multivariante la anemia por deficiencia de folatos fue el diagnóstico con mayor fuerza de asociación de manera independiente (los restantes diagnósticos identificados estuvieron sobre el límite del intervalo de confianza)(Tabla 11). La relación entre deficiencia de folatos y cáncer gástrico es controversial, pues si bien algunos estudios epidemiológicos muestran asociación (97), no es claro si la deficiencia es causa o consecuencia de la enfermedad. Por su papel en el metabolismo de las purinas se ha visto que la deficiencia de folatos podría alterar los mecanismos de reparación del ADN y la metilación, además de jugar un papel importante en la síntesis del ADN, lo cual podría favorecer el desarrollo de neoplasias (97); sin embargo, la deficiencia de folatos y de vitamina B12 con frecuencia coexisten haciendo difícil dilucidar el origen real de la anemia y por tanto del cáncer gástrico asociado (98), y adicionalmente, la atrofia gástrica y la hipoclorhidria, que son parte del proceso de carcinogénesis, reducen la absorción de vitamina B12 igual que la anemia perniciosa.

Una situación similar se presenta con la anemia por deficiencia de hierro. Se ha descrito asociación de la anemia ferropénica con cáncer gástrico, pero el mecanismo y la

asociación causal no son claros (99, 100). A su vez, la deficiencia de hierro es una manifestación frecuente de la gastritis crónica atrófica inducida por sangrado oculto, competencia del *H pylori* por el hierro dietario, e hipoclorhidria (101). De hecho, algunos estudios muestran que al momento del diagnóstico de cáncer gástrico cerca del 60% de los pacientes puede tener anemia (bajo nivel de hemoglobina) y cerca de 40% puede ser diagnosticado con anemia ferropénica, pero solo cerca de 20% reportan historia previa de anemia ferropénica (102).

Los datos sugieren una elevada frecuencia de anemia en los pacientes con cáncer gástrico, pero a su vez una dificultad relativamente importante para hacer una adecuada clasificación y diagnóstico de la anemia, a lo que se suman las limitaciones de la codificación de los registros administrativos utilizados en este estudio. Por ello se decidió agrupar todas las anemias carenciales en la construcción del modelo multivariante, a partir de lo cual se encontró una asociación significativa del diagnóstico de anemia con el cáncer gástrico, la cual se mantuvo en el análisis de tiempo al evento (Tablas 17 y 18, Gráfico 4). No obstante, la métrica mostró un desempeño modesto del modelo predictivo (AUC 0,78 IC95% 0,75-0,80) y consistente con ello, en el análisis mediante aprendizaje de máquinas el diagnóstico de anemia solo se reportó dentro de las variables con mayor capacidad predictiva en el modelo de regresión logística LASSO para el último año (OR 2,4 IC95% 2,2-2,6). Este resultado puede relacionarse con el hecho de que el diagnóstico de anemia se haría en etapa avanzada de la enfermedad, cercano al momento del diagnóstico de cáncer gástrico. Aun así, el análisis de tiempo al evento reportó una mediana de 28,2 meses al diagnóstico de cáncer gástrico en los sujetos con anemia, sugiriendo un intervalo de tiempo suficiente para mejorar el estado clínico al diagnóstico mediante un protocolo de vigilancia endoscópica apropiado, si la selección de pacientes a incluir en dicho programa puede hacerse con mayor precisión.

El consenso colombiano para el diagnóstico temprano del cáncer gástrico sugiere realizar endoscopia en todos los pacientes adultos con síntomas de dispepsia, a fin de identificar las personas a ingresar en el programa de seguimiento endoscópico (45). La prevalencia de dispepsia es difícil de establecer porque depende de la población estudiada (sesgo de selección), de los criterios diagnósticos y de los métodos de recolección de información. Un metanálisis que reunió información global encuentra que la prevalencia en adultos de América Latina varía de 7,0% en Centroamérica a 37,7% en Suramérica (no incluye información de Colombia)(103); sin embargo, un consenso mexicano sobre el tema que reúne diversos estudios de ese país encuentra que la prevalencia utilizando criterios diagnósticos uniformes varía entre 12,0% y 68,0% en las diferentes regiones (104). Además de la elevada variabilidad, la información de estudios poblacionales no sugiere una relación entre la prevalencia de dispepsia y la incidencia de cáncer gástrico, a pesar de su asociación con la infección por H pylori; de hecho, una revisión de estudios poblacionales a nivel global encuentra las prevalencias más altas en Medio Oriente y África (105), esta última la región con la incidencia de cáncer gástrico más baja en el mundo a pesar de la elevada prevalencia de *H pylori*.

Una alternativa para la selección de individuos a seguir mediante endoscopia son las escalas de riesgo, las cuales se construyen a partir de factores de riesgo o de su combinación con síntomas de cáncer gástrico y biomarcadores. De los modelos multivariantes identificados con adecuado desempeño (AUC>0,8), cinco utilizan

parámetros histológicos o endoscópicos como parte de las variables predictivas (atrofia, metaplasia intestinal, estado de la mucosa gástrica)(12, 18, 21, 47, 106); es decir, no son modelos utilizables para la selección de población a incluir en el programa de seguimiento endoscópico sino que predicen la probabilidad de cáncer gástrico para quienes ya se encuentran en el programa. Un modelo adicional desarrollado a partir de la identificación de 12 marcadores de microARN circulante reportó un AUC de 0,85 con sensibilidad 87,0% y especificidad 68,4% (107); pero además de la baja factibilidad de implementación en el cuidado primario, la limitada especificidad supone un elevado número de falsos positivos que podría generar una relación de costo-efectividad no favorable para su implementación en escenarios de ingresos medios y bajos.

Como se anotó en el marco conceptual, las escalas de riesgo se afectan por el riesgo atribuible poblacional, lo que resulta en la necesidad de validación si la escala se aplica en poblaciones distintas a aquella de donde se tomaron los datos para su construcción (108). En la revisión realizada por Fan Z et al, se identificaron 71 modelos predictivos de cáncer gástrico, de los cuales solo ocho se realizaron en poblaciones occidentales (incluyendo Asia Occidental). Aparte de un modelo iraní ya referenciado (106), uno de los modelos está específicamente orientado a cáncer de esófago (109), tres son modelos de pronóstico (metástasis o supervivencia)(110-112), y dos corresponden a estudios de asociación sin desarrollo de un modelo predictivo (113, 114).

El modelo restante se construyó en Reino Unido (19) y utiliza las mismas variables de un modelo previamente desarrollado por Stapley S et al, el cual se propuso para uso en cuidado primario y se basó en síntomas para la selección de pacientes que deben ir a endoscopia a fin de identificar cáncer de esófago o cáncer gástrico (115). El modelo de Stapely utiliza el sexo, la edad, 12 síntomas (incluyendo disfagia y pérdida de peso), y 7 parámetros de laboratorio (pruebas rutinarias). Su evaluación se basó en el VPP sin considerar el AUC, la sensibilidad o la especificidad; en tal sentido la disfagia fue el síntoma con mayor capacidad predictiva (VPP 4,8) y solo la disfagia, dispepsia y vómito se asociaron significativamente con cáncer gástrico 9 meses antes del diagnóstico (máximo periodo de observación). Los autores proponen usar un riesgo de 2% como punto de corte para la selección de los pacientes, con lo cual se requeriría investigar endoscópicamente 55 pacientes por cada caso de cáncer de esófago o estómago identificado. Briggs et al compararon el modelo de Stapley con modelos basados en aprendizaje de máquinas, encontrando para el primero un AUC de 0,81 con sensibilidad 33% y especificidad 91% cuando se asume una probabilidad de detección de 2% como sugieren los autores; utilizando las mismas variables, la regresión logística LASSO y los modelos de máquina de soporte de vectores reportan un mejor desempeño por incremento en la sensibilidad sin reducción de la especificidad (AUC 0,87; sensibilidad 41% y especificidad 91%)(19).

De manera similar, otro modelo de aprendizaje de máquinas desarrollado en Irán (potenciación del gradiente extrema), utiliza variables sociodemográficas (edad, sexo, nivel socioeconómico, área de residencia, índice de hacinamiento), presión arterial, medidas antropométricas (índice de masa corporal, razón cintura/cadera), factores de riesgo asociados al comportamiento (tabaquismo, uso de opioides o alcohol, consumo de bebidas calientes, actividad física), y seis antecedentes patológicos personales (20). El modelo reporta AUC 0,98, sensibilidad 100% y especificidad 99%; sin embargo, en la

matriz de confusión (tabla de 2x2) se muestran 82 de 152 casos de cáncer que no fueron detectados por el modelo lo que indica una sensibilidad de 54%, de forma que la discrepancia entre lo calculado y lo reportado en la matriz de confusión no permite tener claridad frente a los resultados del estudio, los que por demás se acercarían a un desempeño perfecto no publicado antes.

A pesar de la limitada sensibilidad, los modelos de Briggs E et al y Mohammadnezhad K et al reportan una AUC elevada y un alto nivel de especificidad. Además de la factibilidad de aplicación en cuidado primario dada la información que utilizan, el uso de síntomas o antecedentes patológicos como variables predictoras tiene relevancia para efectos de comparación con el presente estudio. El estudio de Briggs et al se basa en síntomas expresados durante el año previo al diagnóstico y consignados en los registros electrónicos de cuidado primario en el Reino Unido (4,8 millones de pacientes en 600 centros de cuidado primario). El estudio Mohammadnezhad K et al incluye antecedentes patológicos basado en información de 15 años de seguimiento en una cohorte del norte de Irán con 16.599 participantes (Golestan Cohort Study)(116), de los cuales se incluyeron 9.842 según los datos reportados en la matriz de confusión. En el presente estudio se utilizaron dos periodos de tiempo para el análisis como son un año y dos años previos al diagnóstico de cáncer gástrico con un periodo de lavado de seis meses para el segundo.

La definición de tiempos de seguimiento reviste gran importancia dada su relación con la probabilidad de detectar la enfermedad en estados tempranos, pero también con los requerimientos para su implementación en la práctica clínica en donde seguimientos más prolongados exigen mayor esfuerzo de los sistemas de salud. Como se anotó en el marco conceptual, entre 40% y 50% de los pacientes con enfermedad temprana podrían manifestar algún síntoma específico en el año previo al diagnóstico, siendo lo más frecuente el dolor abdominal (33), de allí que los modelos que utilizan información del último año muestren buen desempeño; sin embargo, ninguno de los modelos identificados y tampoco los utilizados en este estudio diferenciaron entre estados tempranos y estados avanzados de la enfermedad. En tal sentido, se esperaría que, si no se diferencian estados tempranos y estados avanzados, un seguimiento de dos años incrementaría el número de casos detectados lo que se traduce en mayor sensibilidad (36). No obstante, en este estudio el modelo de regresión logística LASSO a un año mostró un desempeño superior al modelo de los dos últimos años (sensibilidad 81,3% y 68,3% respectivamente), lo cual considero se debe al periodo de lavado de seis meses utilizado para este último periodo de análisis. El periodo de lavado pretende desestimar síntomas y diagnósticos que podrían tener elevado valor predictivo pero que son más comunes en enfermedad avanzada que en enfermedad temprana (en Colombia el 75% de los casos se detectan en estado avanzado)(3), y por ello es posible que los modelos que analizan el periodo dos años atrás del diagnóstico en este estudio se acerquen más al desempeño del modelo para la detección de enfermedad en estados tempranos.

En relación con las variables predictivas, el modelo de Briggs E et al utilizaron edad, sexo, 12 síntomas, y siete pruebas de laboratorio. El modelo de Mohammadnezhad K et al utiliza seis antecedentes patológicos personales e información detallada de 8 factores de riesgo que difícilmente se obtiene por fuera de un estudio como el que soporta la construcción del modelo. El presente estudio utilizó 149 códigos de diagnóstico del CIE10

en el modelo del último año previo al diagnóstico y 173 códigos en los modelos que analizan dos años atrás del diagnóstico, recolectados en la rutina de los servicios de salud del país. La comparación de las variables de síntomas o antecedentes patológicos se presenta en la Tabla 21.

Tabla 21. Síntomas y diagnósticos reportados como variables predictivas en modelos de aprendizaje de máquinas seleccionados

Briggs E et al	Mohammadnezhad K et al	Murillo R (Tesis)
Dolor abdominal	Cardiopatía reumática	Malformación intestinal
Dolor torácico	Cardiopatía coronaria	Alteración del habla
Constipación	Hipertensión	Osteocondrosis juvenil
Tos	Falla renal	Tumor maligno de ovario
Dispepsia	Icterica/Hepatopatía	Gastritis y duodenitis
Disfagia	ACV	Otras enfermedades del estómago
Dolor epigástrico		Dispepsia
Fatiga		Tumor maligno de piel
Náusea/Vómito		Tumor maligno de colon
Reflujo		Influenza
Dificultad para respirar		Anemia carencial
Pérdida de peso		Tumor maligno de encía
		Hernia de cavidad abdominal
		Nefritis crónica
		Síntomas alimentarios

Para el presente estudio (Murillo – Tesis), se presentan las variables con mayor aporte en la categoría de diagnóstico para el modelo que analiza el último año previo al diagnóstico

El modelo de Briggs y el presente estudio utilizaron información poblacional; sin embargo, en el modelo de Briggs los síntomas se seleccionaron a partir del valor predictivo para el diagnóstico de cáncer gástrico reportados previamente en el modelo de Stapley S et al (115). El modelo de Mohammadnezhad utilizó información de un grupo seleccionado de pacientes bajo observación por un periodo de tiempo prolongado; además de su asociación con cáncer gástrico en el estudio de cohorte, los diagnósticos finalmente incluidos como antecedentes en este modelo, fueron seleccionados por un grupo de expertos; en este caso el sesgo de recordación puede jugar un papel importante en el reporte de síntomas, sin embargo, el uso de diagnósticos confirmados en lugar del reporte de síntomas podría reducir el efecto de dicho sesgo.

Es de notar que en ambos modelos (Briggs y Mohammadnezhad), los síntomas o diagnósticos para inclusión como variables predictivas fueron seleccionados a priori, de la misma forma que se hizo en el modelo multivariante del presente estudio y de forma diferente a la aproximación utilizada en los modelos de aprendizaje de máquinas. La inclusión exclusiva de factores de riesgo conocidos introduce el sesgo del investigador en la construcción de los modelos y limita la posibilidad de identificar nuevas variables predictoras. En ese contexto, la mejoría en la capacidad predictiva de un modelo basado en aprendizaje de máquinas estaría dada esencialmente por una mejor asignación de coeficientes a las variables analizadas en modelos epidemiológicos multivariantes, tal como parece ocurrir en el modelo de Briggs et al (19), en donde la regresión logística LASSO y el soporte de máquinas de vectores mostraron mejor desempeño; sin embargo, los otros modelos de aprendizaje de máquinas evaluados no tuvieron un desempeño

global significativamente superior al modelo multivariante. De manera similar, Huang RJ et al comparan un modelo multivariante para la predicción de cáncer gástrico en pacientes en seguimiento endoscópico frente a cuatro modelos de aprendizaje de máquinas sin encontrar diferencias significativas (21).

Desde la historia natural de la enfermedad, se ha descrito la asociación de la infección por *H pylori* con varias enfermedades extra-intestinales; sin embargo, los mecanismos patogénicos de dicha asociación no son claros. Las hipótesis incluyen la inducción de respuesta inflamatoria sistémica, la expresión de proteínas que mimetizan péptidos del huésped y la interferencia en la absorción de nutrientes y medicamentos (117). Estos mecanismos de forma aislada o mediante su interacción, se han propuesto para explicar la asociación encontrada entre la infección por H pylori (particularmente de las cepas más virulentas CagA positivas) y enfermedad autoinmune (118), enfermedad aterosclerótica y sus desenlaces a nivel cardiovascular (119), hipertensión arterial (120), glaucoma (121), rosácea y urticaria crónica (122), y púrpura trombocitopénica (123). Cabe resaltar que la relación causal más allá de la asociación sigue siendo controversial, pues algunos estudios realizados en población asintomática no encuentran elevación de marcadores de inflamación en pacientes con infección por *H pylori* (117), y adicionalmente la mayoría de estas patologías y otras como los trastornos cognitivos (124, 125), han mostrado asociación no solo con infección por *H pylori* sino también con infección por otros microorganismos.

Aparte de la infección por *H pylori* y la anemia ya discutidos, el cáncer gástrico se ha asociado con otras patologías como la diabetes (particularmente cáncer del cardias)(126), lo que en su conjunto puede traducirse en una expresión sintomática compleja en el proceso de carcinogénesis. En el caso de la diabetes, se sugiere que la asociación se explica por una interacción compleja de los factores de riesgo compartidos como la obesidad, hiperglicemia y resistencia a la insulina, una menor tasa de eliminación de la infección por *H pylori* y la coexistencia de morbilidad como los trastornos cardiovasculares (127), pero la búsqueda de explicación a una relación específica ignora la complejidad del proceso y los potenciales efectos sistémicos del desarrollo tumoral.

El avance en el entendimiento de la biología del cáncer ha estado centrado en las características fenotípicas de la célula y el microambiente tumoral, pero el entendimiento de las interacciones del proceso de carcinogénesis con órganos distantes y la fisiopatología de la respuesta del huésped a esta agresión han tenido menos desarrollo (128). De hecho, la investigación en torno a marcadores biológicos para el diagnóstico temprano del cáncer se enfoca prioritariamente en estos componentes del proceso de carcinogénesis (129); sin embargo, en este proceso aspectos como las alteraciones del sistema inmune y del metabolismo van mucho más allá del microambiente tumoral. En su desarrollo, el tumor libera no solo ADN y ARN al torrente sanguíneo, sino además moléculas paracrinas y células con potencial inmunogénico que inducen la producción de moléculas proinflamatorias, las que a su vez podrían inducir alteraciones en el proceso de hematopoyesis. De manera similar, la reprogramación del metabolismo en la célula tumoral induce modificaciones en el metabolismo sistémico que aún no tienen clara comprensión, lo cual puede darse de forma directa por la afectación en la disponibilidad de nutrientes igual que lo descrito para el H pylori, o por la reprogramación metabólica en órganos distantes como el hígado y el cerebro (128).

En esta medida, pretender identificar el proceso de carcinogénesis en etapas tempranas a partir de un único marcador biológico o síntoma podría verse como una iniciativa quimérica. Existe evidencia suficiente de que un antecedente o síntoma aislado no tiene mayor valor predictivo para el diagnóstico de cáncer gástrico, y como se anotó el uso de síntomas con elevado valor predictivo para el diagnóstico de cáncer gástrico no tiene mayor valor clínico en etapas tempranas de la enfermedad. Es posible que la conjunción de múltiples patologías como expresión de factores de riesgo o cambios en el proceso de carcinogénesis permita configurar un perfil de riesgo basado en conglomerados de síntomas, lo que posibilitaría identificar sujetos que deben priorizarse para el seguimiento endoscópico. En tal sentido, los modelos de aprendizaje de máquinas en el estudio identificaron códigos de diagnósticos y procedimientos relacionados no solo con patología gastrointestinal sino además con enfermedades cardiovasculares, trastornos cognitivos, trastornos alimentarios, enfermedad genitourinaria, trastornos oftalmológicos y enfermedad dermatológica (Anexo 4).

La explicación de la asociación de estos diagnósticos con el diagnóstico posterior de cáncer gástrico en este estudio, puede estar en la argumentación expuesta sobre la expresión clínica de los factores de riesgo para cáncer gástrico el efecto sistémico del proceso de carcinogénesis; sin embargo, hay también variables diagnósticas y de procedimiento que podrían ser factores de confusión en el análisis adelantado, dentro de las que llama la atención la presencia de otros tumores malignos en las variables con mayor capacidad predictiva. La incidencia de enfermedad metastásica al estómago es muy baja (130), por lo que los reportes de esta condición usualmente se limitan a casos aislados o pequeñas series de casos. No obstante, prácticamente todas las localizaciones anatómicas identificadas en el presente estudio han sido reportadas como sitios originarios de metástasis al estómago (130, 131), y llama particularmente la atención el tumor maligno de ovario, debido a que la enfermedad de Krukenberg, correspondiente a enfermedad metastásica al ovario con origen mayoritario en el estómago, con frecuencia representa un reto diagnóstico para la identificación del sitio primario (132).

En el tratamiento de datos se hizo una juiciosa revisión de los antecedentes de diagnósticos o procedimientos asociados a cáncer gástrico a fin de eliminar los individuos con tal condición antes del periodo definido para el diagnóstico (septiembre a diciembre de 2017), y mejorar en esa medida la capacidad discriminativa del modelo. Sin embargo, la identificación de la enfermedad metastásica y su errónea clasificación como cáncer primario del estómago escapa a los procedimientos utilizados, y en una base de datos de la magnitud y características descritas para este estudio, aun los eventos de baja frecuencia pueden encontrar una fuerza de asociación elevada.

Los modelos de Briggs y Mohammadnezhad (19, 20), usaron otras variables predictoras para mejorar el desempeño de sus modelos (resultados de laboratorio y factores de riesgo), información que desafortunadamente no está disponible en las bases de datos en las que se fundamenta este estudio. No obstante, en el presente estudio se utilizaron otras variables disponibles incluyendo procedimientos realizados a los pacientes durante el periodo de observación, municipio de residencia y compañía de seguros. Los procedimientos identificados dentro de las variables con mayor aporte predictivo son coherentes con los diagnósticos, pero algunos de ellos toman mayor relevancia en esta

categoría de variables como los procedimientos oftalmológicos (Anexo 4). A su vez, los municipios de residencia aportan al modelo de forma totalmente coincidente con la epidemiología del cáncer gástrico en Colombia, en donde la ubicación en las cordilleras confiere riesgo incrementado y la ubicación en las planicies y particularmente en la Costa Atlántica confiere riesgo reducido (Gráfico 7)(133).

El elevado valor predictivo del aseguramiento tiene difícil interpretación. El sistema de salud ha desarrollado mecanismos para evitar la selección adversa de enfermedades de alto costo incluyendo el cáncer (3); no obstante, tales mecanismos no estaban aún en operación plena para el periodo de análisis (2015-2017), pues si bien para entonces se reportaban las prevalencias de los diferentes tipos de cáncer en la población del régimen contributivo, no existía compensación económica relacionada con dichas prevalencias. Es posible que la concentración de casos por selección adversa explique el mayor riesgo de los pacientes afiliados a una determinada compañía de seguros de salud; sin embargo, la territorialización de las aseguradoras entendida como una distribución desigual de la población de afiliados en el territorio nacional, puede explicar en parte el riesgo mayor si algunas aseguradoras están concentradas en las regiones de mayor riesgo para cáncer gástrico; desafortunadamente no fue posible ubicar información oficial acerca de la presencia de las aseguradoras por departamento a fin de verificar tal hipótesis. En ese orden de ideas podría pensarse que su asociación es espuria y depende de la ubicación geográfica del paciente (municipio), pero su presencia en varios municipios de la región hace que su valor predictivo sea mayor que el del municipio de forma aislada; de hecho, el aporte de esta variable al valor predictivo del modelo se corroboró mediante un análisis de sensibilidad en el que se corrió el modelo sin la variable, con lo que se observó una desmejora significativa del desempeño (AUC 0,76, sensibilidad 61,2 y especificidad 77,8 con punto de corte 0,5).

En general, los modelos de aprendizaje de máquinas desarrollados en este estudio tuvieron un desempeño similar o superior a los encontrados en la literatura, no solo por el valor del AUC sino además por su balance entre sensibilidad y especificidad, siendo el de mejor desempeño la regresión logística LASSO con periodo de observación un año antes del diagnóstico (AUC 0,92, sensibilidad 70,8% y especificidad 92,0% - Tabla 20). Los modelos con periodo de observación dos años antes del diagnóstico tuvieron menor sensibilidad siendo de nuevo la regresión logística LASSO el de mejor desempeño sin mayor diferencia con los bosques aleatorios y la potenciación del gradiente.

Las diferencias en el desempeño global de los modelos pueden estar relacionadas con la parametrización (ver sección de métodos), por lo que una calibración de parámetros más extendida podría llevar a resultados más cercanos entre modelos (68); sin embargo, se observaron notables diferencias en el coeficiente asignado a las variables con mayor capacidad predictiva (Gráfico 7). La aseguradora, los síntomas gastrointestinales y los municipios de cordillera fueron en general prevalentes en todos los modelos, pero otras variables diagnósticas y procedimientos tuvieron mayor variabilidad. Las diferencias en las variables predictoras han sido descritas en modelos que usaron un número limitado de predictores (19), podría pensarse que los modelos que utilizan un número extendido de variables como en el presente estudio, tienen mayor tendencia a reportar tal variabilidad, en donde la selección aleatoria de diferentes conglomerados de factores predictores puede converger en la misma o similar capacidad discriminativa.

Implicaciones potenciales para la práctica clínica

La selección de pacientes para seguimiento endoscópico representa en el momento la alternativa más viable para mejorar la detección temprana de cáncer gástrico en países de ingresos bajos y medios y en países de riesgo medio-bajo. Las opciones disponibles para la selección de pacientes incluyen las escalas construidas a partir de factores de riesgo, la identificación temprana de síntomas, el uso de pruebas de laboratorio implementables en cuidado primario, o la combinación de ellas.

En este estudio se abordó la expresión sintomática de la enfermedad traducida en diagnósticos clínicos observados durante el periodo de latencia del cáncer gástrico. Los síntomas de alarma como disfagia, hematemesis o pérdida de peso tienen un VPP suficientemente alto para justificar la endoscopia lo cual ya hace parte de la práctica clínica actual (115); sin embargo, el uso exclusivo de síntomas de alarma para la selección de pacientes tiene baja sensibilidad, lo cual es aún más crítico en estados tempranos de la enfermedad dada su baja prevalencia durante esta etapa (19, 33, 34).

Una alternativa para mejorar el VPP de los modelos desarrollados puede ser una mejor selección de la población objeto para su implementación. Como se anotó, en Colombia la incidencia de cáncer gástrico tiene un rango que va desde 2,5 hasta 29,7 por 100.000, por lo que el uso de los modelos tendría mejor desempeño en poblaciones de alto riesgo ya identificadas por los estudios epidemiológicos en Colombia, lo cual se refuerza con los hallazgos referentes al valor predictivo de los municipios y las aseguradoras en salud. Adicionalmente, el uso de los modelos desarrollados en un rango de edad más restringido podría también contribuir a mejorar su desempeño. En este sentido, los modelos de regresión logística LASSO mostraron el mejor desempeño y sus resultados fueron más coherentes con la tendencia por edad de la enfermedad, en la que Colombia alcanza tasas de incidencia por encima de 20 por 100.000 solo a partir de los 50 años, lo que sugiere que poblaciones menores de esta edad podrían no ser priorizadas.

Se ha propuesto hacer una mejor selección y combinación de los síntomas a utilizar en los modelos predictivos a fin de aumentar el VPP (115); sin embargo, esta alternativa llevaría invariablemente a una mayor reducción de la sensibilidad. Por ello la búsqueda de un elevado VPP en la expresión de síntomas debería revisarse a la luz de la alternativa de tamización poblacional o la selección de personas asintomáticas a partir de escalas de riesgo, las cuales implican la inclusión del 50% al 100% de la población en el programa de seguimiento endoscópico.

En la práctica es poco probable que una sensibilidad adecuada se logre sin una alta tasa de falsos positivos. A fin de mejorar el desempeño de su modelo, Stapley et al proponen bajar el umbral de predicción de 2% a 1% (115); pero aun así, la sensibilidad alcanzada es solo 41% (19). En el presente estudio se obtuvieron VPP menores de 1% en todos los modelos, lo que indica una elevada tasa de falsos positivos; sin embargo, el umbral de predicción para la mayor AUC es alto (60%), lo que se traduce en la detección de la mayoría de casos con el seguimiento de una fracción menor de la población; específicamente se requiere seguir al 10,1% o 16,1% de la población para detectar el 70% de los casos de cáncer gástrico en el seguimiento de un año y dos años, lo que representa un número necesario a seguir de aproximadamente 790 y 1170, respectivamente.

En ambos casos la proporción de población a seguir es sustancialmente menor que la requerida para la tamización poblacional o la selección de individuos asintomáticos a partir de escalas de riesgo. El número necesario a seguir es significativamente mayor que el reportado para la reducción de mortalidad con la tamización de otros tipos de cáncer como el de mama (90 a 257)(134); sin embargo, tales números se obtienen con un seguimiento de 20 a 25 años y datos reportados en tamización de cáncer gástrico en China sugieren que dependiendo del riesgo poblacional el número necesario a tamizar para detectar un caso de cáncer en la población de 50 a 64 años oscila entre 370 y 1860 con seguimiento a un año, y entre 490 y 2586 con seguimiento a cinco años (135).

Limitaciones del estudio

La principal fortaleza del trabajo es el uso de una gran base de datos que tiene las ventajas discutidas con anterioridad (base poblacional, seguimiento prospectivo, no selección a priori de variables, estudio de enfermedades de baja prevalencia, análisis de múltiples dominios)(54); sin embargo, el origen de la información en una base de datos administrativa constituye una de las principales limitaciones del estudio. Se desarrolló un complejo algoritmo para la verificación del etiquetado de cáncer gástrico como desenlace y de la anemia como variable predictora en el modelo multivariante; sin embargo, aparte de estas dos condiciones patológicas el estudio utilizó como variables independientes entre 150 y 170 códigos diagnósticos y entre 120 y 150 códigos de procedimientos, ninguno de los cuales tuvo verificación de las etiquetas. A pesar de que ambas codificaciones constituyen la base del registro de información en el sistema de salud lo cual es una fortaleza, la heterogeneidad y el error en el uso de los códigos son elevados por lo que una misma condición puede registrarse de diversas formas o un código puede con elevada frecuencia no corresponder a la condición patológica diagnosticada. Esta situación reduce la capacidad predictiva del modelo y genera dificultad en la interpretación de los resultados, por lo que utilizar un modelo de aprendizaje supervisado y de caja abierta resulta conveniente para el uso de este tipo de datos, de forma que se pueda hacer una limpieza progresiva de la información, de manera que este proceso resulta fundamental en la construcción de los modelos predictivos y de hecho tomó una parte considerable del tiempo invertido en desarrollo del proyecto.

El ejercicio de limpieza de datos (reducción de la dimensionalidad) puede sin embargo llevar a un sobreajuste del modelo, situación que podría estar presente en los modelos de bosques aleatorios y potenciación del gradiente, en los que a pesar de su cercanía, el AUC del entrenamiento fue superior al AUC de prueba. Esta situación podría explicar el menor desempeño frente a los modelos de regresión logística LASSO. El elevado costo computacional exigió en este ejercicio correr estos dos modelos sin sintonización, proceso que debe adelantarse a futuro si se quiere mejorar su desempeño.

La búsqueda de un mejor modelo podría hacerse a futuro utilizando la potenciación del gradiente extremo (XGBoost por su denominación en inglés). XGBoost utiliza el mismo principio de selección progresiva del mejor modelo al aprender de cada versión y corregir el error como lo hace la potenciación de gradiente, pero incluye algunas características adicionales (67). En este trabajo se hizo un ejercicio de regularización con LASSO previo a la utilización de los modelos de bosques aleatorios y potenciación del gradiente; sin embargo, XGBoost incluye la regularización no solo mediante la técnica LASSO sino también mediante la técnica *ridge* lo que permite al mismo tiempo hacer una mejor

selección de variables y reducir la colinealidad. Desafortunadamente, los recursos computacionales disponibles en desarrollo del trabajo no nos permitieron usar este modelo en el actual ejercicio de investigación.

En el estudio las limitaciones de la base de datos tuvieron mayor efecto en la construcción del modelo multivariante no solo por las deficiencias en la calidad de la información, la cual es más relevante cuando se usa un número reducido de variables, tal como ocurrió en este caso debido a la disponibilidad de un número limitado de categorías de variable. Además de los síntomas y antecedentes personales, los trabajos de Briggs E et al y Mohammadnezhad K et al utilizaron 7 pruebas de laboratorio y 8 factores de riesgo respectivamente (19, 20). En la revisión de Fan Z et al (17) se identificaron tres modelos mas, desarrollados en poblaciones de Asia Oriental y apropiados para la selección de sujetos a incluir en seguimiento endoscópico, los cuales utilizaron de manera común la edad y la historia familiar de cáncer gástrico; al menos dos de ellos utilizan conjuntamente el sexo, el IMC, el consumo de frutas y verduras y el consumo de tabaco y alcohol; pero uno utiliza además marcadores de atrofia (pepsinógeno, *H pylori*, G-17)(136-138). Como se observa, todos los modelos multivariantes encontrados utilizan un numero de categorías más amplio incluyendo pruebas de laboratorio, factores de riesgo y antecedentes familiares, mientras que dada la disponibilidad de información en este estudio solo se utilizaron, además de la anemia, 17 antecedentes patológicos personales reunidos en el índice de comorbilidad de Charlson, la edad y el sexo.

La heterogeneidad en los modelos multivariantes y en los modelos de aprendizaje de máquinas resalta su dependencia de la disponibilidad de información y su uso restringido a las poblaciones en donde se desarrollan. Es posible entonces que el uso más extendido de datos en el modelo multivariante desarrollado resulte en un mejor desempeño. En particular, agregar información sobre el estado inflamatorio de bajo grado asociado a *H pylori* podría representar una mejora significativa de la capacidad predictiva, lo cual se podría hacer con resultados de laboratorio de cuidado primario como el recuento de leucocitos y plaquetas, la medición de proteína C reactiva, la velocidad de sedimentación globular, la relación neutrófilos/linfocitos o la relación plaquetas/linfocitos (139). Desafortunadamente la naturaleza administrativa de la base de datos utilizada no contempla este tipo de información y en la revisión de literatura no se encontró una fuente potencial de este tipo de información en el país.

Ya se mencionó previamente la limitación impuesta por la imposibilidad de identificar los pacientes a quienes se les diagnosticó la enfermedad en estado temprano y los pacientes con enfermedad metastásica al estómago. Tal ejercicio representa un reto mayor dadas las características y las deficiencias en la calidad de la información anotadas; sin embargo, podría intentarse en el caso del estado clínico hacer una combinación de códigos con los procedimientos exclusivos de la enfermedad temprana como la cirugía; es decir, se podría asumir un sistema de etiquetado en el que los casos incidentes de cáncer gástrico (CIE10 mas código CUPS de biopsia), se asumen como enfermedad temprana si fueron llevados a gastrectomía (código CUPS). La valoración de tal abordaje requiere un nuevo ejercicio de investigación.

Uno de los principales retos del aprendizaje de máquinas en salud es del desbalance de las bases de datos, lo que significa que usualmente los pacientes con la enfermedad

bajo estudio constituyen una clase minoritaria frente a los pacientes sin la enfermedad (140). El desbalance es en realidad la situación más común en la vida real; sin embargo, los modelos de aprendizaje han sido diseñados asumiendo una distribución similar entre clases por lo que tienen un sesgo innato hacia los grupos mayoritarios favoreciendo una mejor clasificación del desenlace en este grupo comparado con la clase minoritaria. Los métodos para mejorar el balance de los datos comprenden la manipulación de los datos, la modificación de los algoritmos de aprendizaje o la combinación de los dos. En el primer caso se puede hacer un sobre-muestreo de la clase minoritaria o un sub-muestreo de la clase mayoritaria y en el segundo caso se pueden usar funciones automatizadas o la parametrización manual de los algoritmos. En el presente estudio se utilizaron las dos técnicas, en primer lugar, reduciendo los registros con códigos diagnósticos y de procedimientos con poca probabilidad de asociación con el cáncer gástrico (causas externas, embarazo, etc.), lo que resultó en la eliminación mayoritaria de sujetos sin la enfermedad y en segundo lugar se utilizaron parámetros automatizados disponibles en los algoritmos de aprendizaje.

Los resultados muestran que los modelos dan mayor peso a la especificidad; no obstante, el VPP fue inferior a 1% lo cual puede verse como una limitación mayor. Debe tenerse en cuenta que no se utilizó un diseño de tipo corte-transversal o estudio de prevalencia y adicionalmente el ajuste del balanceo previamente descrito pudo haber afectado este desenlace en particular, por ello la interpretación del VPP debe hacerse con cautela. Steinberg DM et al sugieren hacer un ajuste del VPP basado en la prevalencia real de la condición estudiada en la población de estudio cuando el cálculo de este desenlace no proviene de estudios de prevalencia (corte-transversal)(141); desafortunadamente no fue posible para el presente ejercicio encontrar información de prevalencia de cáncer gástrico en la población colombiana de 30 a 89 años perteneciente al régimen contributivo para el año 2017 o aledaños por lo que tal ajuste no pudo ser realizado. Adicionalmente, el número necesario a seguir y el tiempo corto de seguimiento (dos años), sugieren un buen desempeño del modelo a pesar del bajo valor de VPP, esto si se toma en cuenta que valores mayores de VPP en estudios de tamizaje se obtienen tras seguimientos de diez años o mas, los que además incluyen la totalidad de la población.

Debe mencionarse que el objetivo secundario relativo a la construcción de un modelo para cáncer de páncreas como enfermedad con espectro sintomático diferente del cáncer gástrico no pudo ser desarrollado a cabalidad, debido a la complejidad de desarrollar un proyecto paralelo de igual magnitud al que exigía el objetivo primario. Desafortunadamente, esta circunstancia no fue visualizada al inicio del proyecto sino una vez iniciado el trabajo de campo cuando se tuvo mayor información de la dificultad en el manejo de bases de datos de gran tamaño y los retos que implicaba el manejo de datos administrativos. En razón a estas limitaciones, el análisis de cáncer de páncreas solo se hizo de forma exploratoria; no obstante, al ser un objetivo secundario no afecta el cumplimiento del objetivo primario del proyecto. De cualquier forma, los resultados de este ejercicio presentados en forma de anexo, contribuyeron a la selección de modelos de aprendizaje de máquinas para cumplimiento del objetivo primario.

10. Conclusiones

El cáncer gástrico representa la primera causa de muerte por cáncer en Colombia, a pesar de lo cual, no se evidencian esfuerzos del sistema de salud para el control de la enfermedad.

Los consensos internacionales proponen un abordaje diferencial para la detección temprana del cáncer gástrico dependiendo del riesgo poblacional; sin embargo, no existe evidencia clínica de alto nivel que sustente tales propuestas. En este contexto, la selección de pacientes para vigilancia endoscópica representa la alternativa más apropiada para países de ingresos medios y bajos como en América Latina.

A pesar de la carga de enfermedad y de la experiencia en la investigación epidemiológica del cáncer gástrico, la actividad investigativa sobre alternativas para la detección temprana es precaria en la región latinoamericana. El aprendizaje de máquinas brinda nuevas oportunidades de investigación utilizando información disponible en nuestro medio, la cual difiere sustancialmente de la información aportada en regiones del ato riesgo como los países asiáticos o en países de altos ingresos como los de Europa Oriental. No obstante, no se encontraron antecedentes de estudios regionales que utilicen esta herramienta para mejorar el uso de información y la generación de conocimiento sobre el tema.

Dos resultados relevantes se derivan del presente trabajo: el abordaje conceptual utilizado para el uso de las metodologías de aprendizaje de máquinas y la potencial aplicación clínica de los hallazgos.

Desde la perspectiva conceptual este trabajo aborda el proceso de instauración de la enfermedad asumiendo como hipótesis de trabajo los efectos sistémicos del principal factor de riesgo para la enfermedad (infección por H pylori) y el carácter sistémico del proceso de carcinogénesis, el cual usualmente se asume como un fenómeno de predominancia local en el órgano afectado.

En tal sentido, al utilizar toda la información clínica en la fase previa al diagnóstico de la enfermedad, los modelos de aprendizaje de máquinas desarrollados en el presente estudio constituyen un abordaje innovador ya que los ejercicios previos aplicables a la selección de pacientes para vigilancia endoscópica se han limitado a mejorar la capacidad predictiva de modelos epidemiológicos pre-existentes utilizando únicamente variables con conocida asociación con el cáncer gástrico.

Adicionalmente, la inclusión de todas las variables disponibles en las bases de datos del sistema de salud utilizada termina conjugando información clínica (manifestación local y posible manifestación sistémica de la enfermedad), con el efecto de determinantes biológicos (edad, sexo), de determinantes ambientales (municipio), y de determinantes del sistema de salud (compañía de seguros), en un análisis de interrelaciones que no sería posible con los métodos epidemiológicos tradicionales.

Este abordaje metodológico podría ser extensivo al estudio de otras enfermedades con elevada letalidad, incluyendo otras formas de cáncer, a fin de mejorar su diagnóstico temprano. Adicionalmente, la metodología puede hacerse extensiva a otros escenarios en países de ingresos medios en donde la disponibilidad de grandes bases de datos que combinen patrones biológicos con información clínica es escasa, pero en los que es

factible encontrar información sobre diagnósticos y procedimientos asociados a las atenciones en salud, lo cual tiene tendencia creciente en la medida en que se fortalecen las bases de datos administrativos de los sistemas de salud.

En relación con la aplicación clínica de los resultados, a pesar de su menor desempeño predictivo, la asociación encontrada entre el diagnóstico de anemia y el posterior diagnóstico de cáncer gástrico en el modelo multivariante sugiere la pertinencia de considerar la vigilancia endoscópica en pacientes que no tienen causa establecida y no resuelven el cuadro de anemia en un periodo breve de tiempo.

Un elemento adicional por destacar es la identificación del potencial predictivo de cada variable utilizada en los modelos de aprendizaje de máquinas, lo cual posibilita su uso en la construcción de algoritmos para la clasificación de riesgo de los pacientes y así decidir su inclusión en un programa de seguimiento clínico. Tal característica contrasta con modelos de caja negra en donde no es posible identificar el potencial predictivo de las variables incluidas.

Al respecto, en este estudio la regresión logística LASSO mostró mejor desempeño que el modelo multivariante, el modelo de bosques aleatorios y el modelo de potenciación del gradiente, debido en el caso del modelo multivariante a la limitada disponibilidad de información y en el caso de los modelos de aprendizaje de máquinas a una parametrización que podría mejorarse mediante sintonización adecuada; sin embargo, la superioridad de la regresión logística LASSO es coherente con reportes previos (19). El desempeño reportado por la regresión logística LASSO en el presente estudio sugiere un importante potencial para su uso en la práctica clínica dada su capacidad discriminativa y el número de pacientes necesario a seguir para detectar un caso de cáncer gástrico. El desempeño observado puede de hecho mejorarse si su implementación se hace en la población de mayor riesgo en el país como son los residentes en los municipios de alta montaña y los mayores de 50 años; sin embargo, reducir el efecto negativo de las limitaciones en el actual ejercicio de investigación amerita nuevos estudios. De igual forma, el uso de los modelos en la rutina de los servicios debe estar precedido de su correcta validación clínica.

11. Agradecimientos

Agradecimiento a Daniel Santiago Bonilla y Lina Buitrago su apoyo en el análisis de datos.

Agradezco a la doctora Claudia Uribe, directora del Registro Poblacional de Cáncer de Bucaramanga por su apoyo con los datos del registro que aportaron a la validación de etiquetas de cáncer gástrico.

Agradezco al Hospital Universitario San Ignacio y a la Pontificia Universidad Javeriana por el apoyo brindado durante el programa académico del doctorado. La disponibilidad de tiempo y el acceso a recursos institucionales que me permitieron hicieron posible el cumplimiento de los objetivos académicos.

Agradezco al comité tutorial de tesis integrado por los doctores Fernando de la Hoz y Mario Góngora, por su retroalimentación y aportes a lo largo del desarrollo del proyecto

De manera especial agradezco a la doctora Martha Serrano por la motivación que me

brindó desde siempre para cumplir este sueño, por su permanente acompañamiento y por su permanente orientación en los asuntos académicos.

Finalmente agradezco al cuerpo docente del Doctorado en Oncología de la Universidad Nacional por haberme permitido seguir aprendiendo y creciendo en mi vida profesional.

12. Anexos

Anexo 1. Variables adicionales identificadas en el proceso de selección para el modelo multivariante

Variable	Tipo de variable	OR	IC95%	
			LI	LS
Manejo de síntomas asociados a la enfermedad				
Metoclopramida	Medicamento	1,95	1,37	2,55
Omeprazol	Medicamento	1,14	1,00	1,39
Sucralfato	Medicamento	1,04	1,00	1,26
Hidróxido de aluminio	Medicamento	1,01	1,00	1,06
Abordaje diagnóstico de la enfermedad				
Imágenes				
Radiografía de tórax*	Procedimiento	1,31	1,01	1,64
Radiografía tejidos blandos del cuello	Procedimiento	1,99	1,00	3,94
Ultrasonografía endoscópica	Procedimiento	1,03	1,00	1,27
Ultrasonografía de abdomen	Procedimiento	1,95	1,22	2,64
Gamagrafía ósea	Procedimiento	1,90	1,00	3,14
Endoscopia				
Colonoscopia	Procedimiento	1,57	1,20	2,21
Esofagogastroduodenoscopia	Procedimiento	6,44	4,35	8,85
Marcadores				
Antígeno carcinoembrionario	Procedimiento	2,78	2,01	3,57
Fosfatasa alcalina	Procedimiento	1,99	1,24	2,99
Transaminasas (ALT)	Procedimiento	1,50	1,00	2,48
Transaminasas (AST)	Procedimiento	1,40	1,00	2,67
Creatininina	Procedimiento	1,77	1,30	2,60
Citopatología				
Citología de órgano o tejido	Procedimiento	1,01	1,00	1,06

Anexo 2. Métrica de los modelos de aprendizaje de máquinas en el escenario de prueba

Regresión LASSO un año										
Punto de corte	Modelo 1		Modelo 2		Modelo 3		Modelo 4		Modelo 5	
	Sensibilidad	Especifidad	Sensibilidad	Especifidad	Sensibilidad	Especifidad	Sensibilidad	Especifidad	Sensibilidad	Especifidad
0,1	100,00%	22,99%	99,45%	23,47%	100,00%	24,60%	99,49%	23,88%	100,00%	22,49%
0,2	93,79%	57,47%	93,99%	60,93%	91,63%	61,33%	93,94%	63,20%	95,83%	59,00%
0,3	86,44%	77,87%	84,15%	78,46%	84,24%	78,71%	82,32%	79,07%	89,58%	77,96%
0,4	80,79%	83,94%	78,14%	83,58%	78,82%	84,19%	75,25%	84,14%	85,42%	83,56%
0,5	75,14%	86,81%	73,22%	86,46%	73,40%	86,91%	70,71%	87,10%	81,25%	86,57%
0,6	67,23%	89,61%	71,58%	89,59%	67,98%	89,68%	62,63%	90,29%	77,08%	89,43%
0,7	61,02%	92,10%	67,21%	92,13%	61,08%	92,54%	55,56%	92,45%	70,83%	92,00%
0,8	53,11%	94,93%	55,74%	95,39%	47,29%	95,63%	42,42%	95,61%	61,46%	95,15%
0,9	32,77%	98,72%	36,07%	98,66%	27,59%	98,85%	26,77%	98,77%	40,10%	98,77%

Regresión LASSO dos años										
Punto de corte	Modelo 1		Modelo 2		Modelo 3		Modelo 4		Modelo 5	
	Sensibilidad	Especifidad	Sensibilidad	Especifidad	Sensibilidad	Especifidad	Sensibilidad	Especifidad	Sensibilidad	Especifidad
0,1	99,55%	10,10%	100,00%	7,60%	99,00%	8,95%	100,00%	7,82%	98,97%	9,32%
0,2	95,48%	35,40%	95,88%	33,31%	96,52%	31,48%	98,02%	31,45%	96,41%	32,22%
0,3	86,43%	62,82%	87,63%	60,85%	88,56%	59,87%	89,60%	59,31%	90,26%	59,05%
0,4	71,49%	81,10%	73,71%	79,81%	75,62%	80,19%	75,74%	79,96%	77,95%	79,38%
0,5	63,35%	86,34%	68,04%	86,01%	68,66%	86,41%	68,32%	85,82%	69,23%	85,94%
0,6	57,92%	89,54%	61,86%	89,18%	61,19%	89,39%	65,35%	89,22%	60,51%	89,70%
0,7	51,58%	92,44%	58,25%	91,92%	54,23%	92,33%	58,91%	92,51%	53,33%	92,49%
0,8	45,25%	94,57%	53,61%	94,50%	48,26%	94,26%	49,01%	94,41%	46,67%	94,61%
0,9	18,10%	99,04%	16,49%	99,07%	18,41%	99,09%	24,26%	99,07%	22,56%	99,05%

Bosques aleatorios 2 años										
Punto de corte	Modelo 1		Modelo 2		Modelo 3		Modelo 4		Modelo 5	
	Sensibilidad	Especifidad	Sensibilidad	Especifidad	Sensibilidad	Especifidad	Sensibilidad	Especifidad	Sensibilidad	Especifidad
0,1	100,00%	0,00%	100,00%	0,00%	100,00%	0,00%	100,00%	0,00%	100,00%	0,00%
0,2	100,00%	0,00%	100,00%	0,00%	100,00%	0,00%	100,00%	0,00%	100,00%	0,00%
0,3	100,00%	0,00%	100,00%	0,00%	100,00%	0,00%	100,00%	0,00%	100,00%	0,00%
0,4	95,78%	28,47%	94,75%	25,84%	95,51%	25,05%	97,45%	18,47%	95,53%	22,87%
0,5	64,61%	87,08%	63,93%	86,67%	68,91%	86,72%	71,27%	83,52%	69,01%	86,69%
0,6	24,68%	98,43%	23,93%	98,60%	21,15%	98,75%	27,64%	98,23%	27,80%	98,75%
0,7	2,60%	99,98%	1,64%	99,97%	1,92%	99,99%	1,45%	99,96%	0,64%	99,99%
0,8	0,00%	100,00%	0,00%	100,00%	0,00%	100,00%	0,00%	100,00%	0,00%	100,00%
0,9	0,00%	100,00%	0,00%	100,00%	0,00%	100,00%	0,00%	100,00%	0,00%	100,00%

Potenciación del gradiente 2 años										
Punto de corte	Modelo 1		Modelo 2		Modelo 3		Modelo 4		Modelo 5	
	Sensibilidad	Especifidad	Sensibilidad	Especifidad	Sensibilidad	Especifidad	Sensibilidad	Especifidad	Sensibilidad	Especifidad
0,1	97,95%	13,82%	99,05%	7,17%	98,65%	9,46%	96,77%	14,58%	98,32%	16,18%
0,2	90,10%	39,40%	87,34%	46,51%	86,82%	41,60%	90,00%	43,69%	89,90%	44,85%
0,3	82,94%	62,33%	72,78%	68,84%	79,05%	65,55%	81,61%	63,21%	79,80%	67,05%
0,4	72,35%	78,76%	66,46%	81,32%	70,61%	79,42%	72,90%	80,36%	67,00%	82,28%
0,5	65,53%	86,12%	61,71%	87,29%	61,15%	87,08%	67,10%	85,89%	62,63%	86,43%
0,6	57,68%	90,81%	57,91%	89,32%	56,76%	89,42%	58,06%	90,28%	57,91%	89,83%
0,7	54,61%	92,17%	52,53%	91,81%	54,05%	91,54%	50,00%	92,87%	52,19%	91,98%
0,8	48,46%	93,76%	47,15%	94,39%	47,64%	93,74%	44,84%	94,26%	42,76%	94,42%
0,9	10,58%	99,52%	8,54%	99,62%	12,50%	99,58%	9,68%	99,69%	10,44%	99,62%

Anexo 3. Área bajo la curva de los modelos de aprendizaje de máquinas en el escenario de prueba

Regresión LASSO 1 año

	Modelo: 7776696	Modelo: 12258477	Modelo: 47586133	Modelo: 76660827	Modelo: 94749095
avgTrain	0,89702727	0,89625345	0,89950746	0,90398896	0,895443149
AUC_test	0,89310763	0,89481628	0,89070256	0,88557914	0,917075474

Tam muestra	10
Confianza	0,95
Alfa	0,05
z	1,95996398

Regresión LASSO 2 años

	Modelo: 391829	Modelo: 25018672	Modelo: 40876102	Modelo: 53537977	Modelo: 64689167
avgTrain	0,85397856	0,84498215	0,8450425	0,843792032	0,849475642
AUC_test	0,84057307	0,85207407	0,85192336	0,865052975	0,856846115

Tam muestra	10
Confianza	0,95
Alfa	0,05
z	1,95996398

Bosques aleatorios 2 años

	Modelo: 10399276	Modelo: 11287013	Modelo: 23144866	Modelo: 39744514	Modelo: 46322612
AUC_train	0,90205261	0,90389065	0,90647204	0,90528295	0,915827853
AUC_test	0,83335327	0,82619835	0,83971155	0,83480201	0,8450267022

maxDepth	5
maxBins	32
numTrees	2000

Potenciación del gradiente 2 años

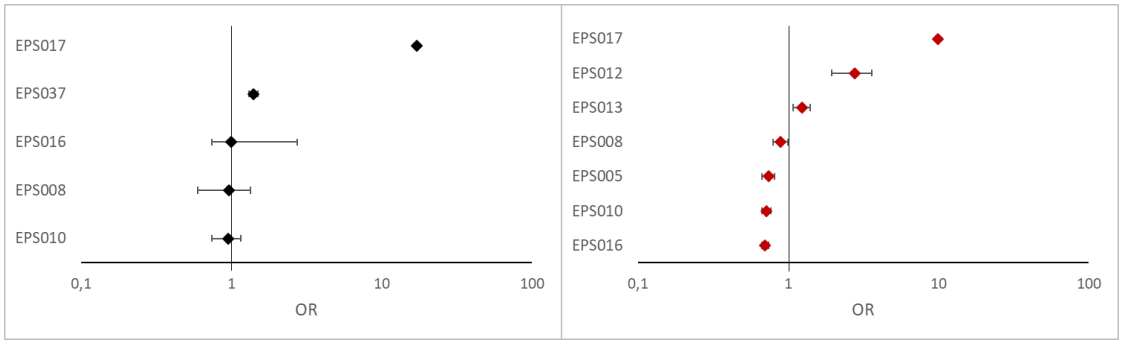
	Modelo: 929035	Modelo: 49601229	Modelo: 56938053	Modelo: 97559788	Modelo: 98862193
AUC_train	0,90205261	0,90389065	0,90647204	0,90528295	0,915827853
AUC_test	0,82169073	0,80415689	0,80629832	0,82085298	0,8219668931

maxDepth	5
maxBins	32
stepSize	0,1

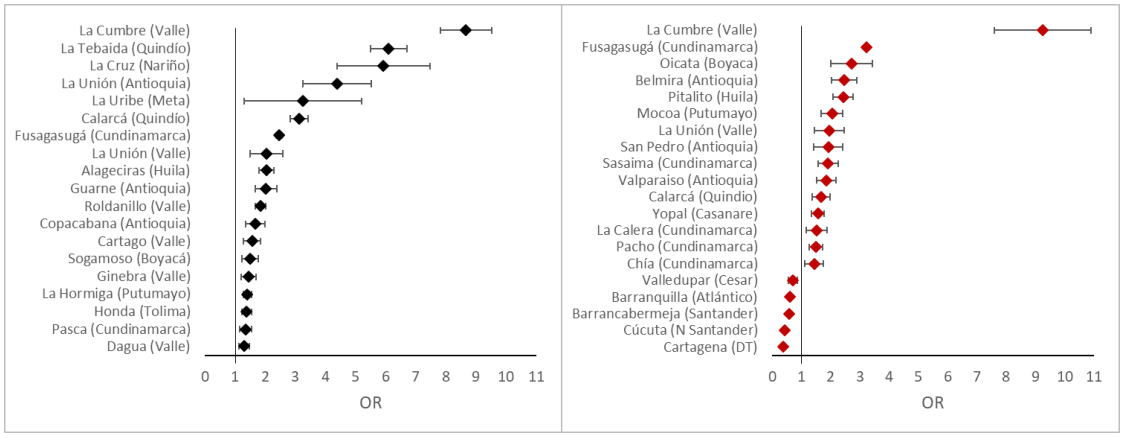
Anexo 4. Contribución de las variables según categoría en los modelos de regresión logística LASSO.

Se presentan las variables con asociación significativa (OR mayor o menor de 1). Los marcadores negros corresponden al análisis del último año y los marcadores rojos al análisis dos años atrás

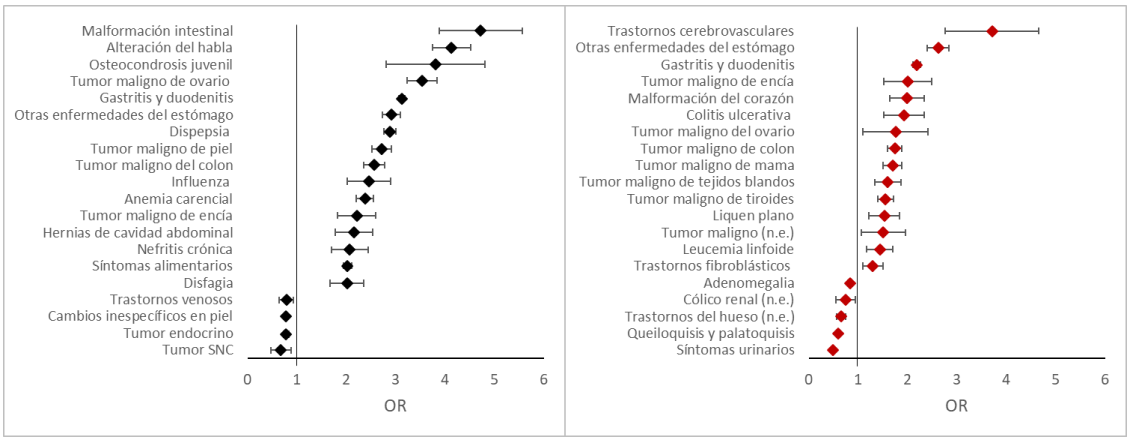
Aseguradoras



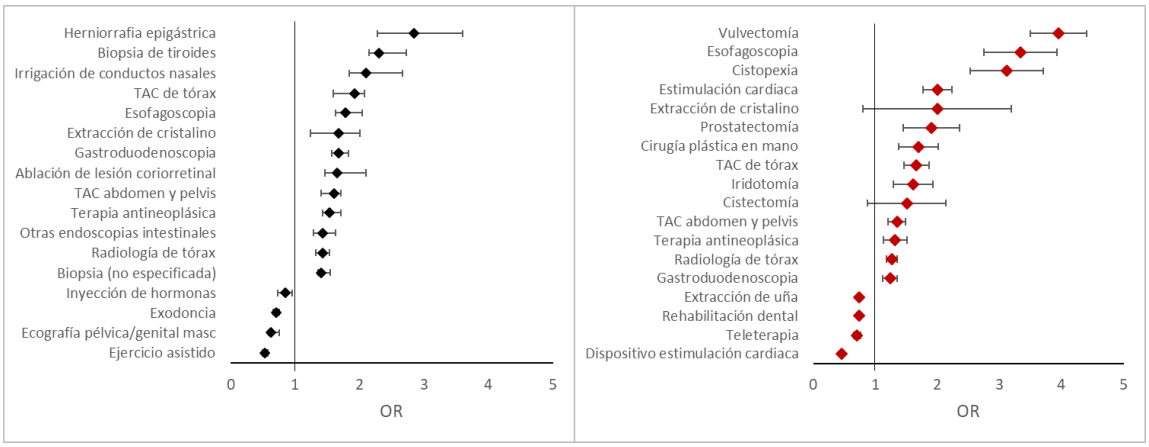
Municipios



Diagnósticos



Procedimientos



Anexo 5. Población a seguir y número necesario a tratar en los modelos de aprendizaje de máquinas

LASSO 1 año

Punto de corte (probabilidad de SI tener cáncer)	Porcentaje de población a intervenir	Porcentaje de pacientes con cáncer intervenidos	Número de personas a seguir
1,000	0,000	0,000	#¡DIV/0!
0,990	0,001	0,029	105,179
0,980	0,001	0,054	144,333
0,970	0,002	0,096	132,793
0,960	0,003	0,131	138,720
0,950	0,005	0,153	157,226
0,940	0,006	0,189	162,661
0,930	0,007	0,229	167,152
0,920	0,009	0,270	172,762
0,910	0,011	0,298	187,594
0,900	0,013	0,327	200,423
0,890	0,015	0,352	217,410
0,880	0,017	0,368	242,874
0,870	0,020	0,385	265,707
0,860	0,023	0,409	288,059
0,850	0,026	0,436	311,671
0,840	0,029	0,456	337,030
0,830	0,033	0,473	367,225
0,820	0,038	0,488	401,666
0,810	0,042	0,510	433,614
0,800	0,047	0,520	468,822
0,790	0,051	0,540	492,041
0,780	0,056	0,557	525,677
0,770	0,060	0,574	547,165
0,760	0,065	0,589	574,886
0,750	0,067	0,596	587,229
0,740	0,069	0,607	594,384
0,730	0,071	0,611	609,981
0,720	0,074	0,619	622,997
0,710	0,076	0,622	635,463
0,700	0,078	0,631	641,403
0,690	0,080	0,640	648,826
0,680	0,082	0,645	665,210
0,670	0,084	0,651	673,818
0,660	0,087	0,659	684,804
0,650	0,089	0,665	697,493
0,640	0,091	0,671	707,700
0,630	0,094	0,675	723,275
0,620	0,097	0,680	741,145
0,610	0,100	0,688	757,954
0,600	0,103	0,693	773,997
0,590	0,105	0,698	786,612
0,580	0,108	0,707	798,037
0,570	0,111	0,710	816,392
0,560	0,116	0,722	833,674
0,550	0,119	0,727	850,822
0,540	0,122	0,731	868,711
0,530	0,125	0,735	884,310
0,520	0,127	0,738	896,422
0,510	0,130	0,744	908,127
0,500	0,132	0,747	923,149
0,490	0,135	0,753	934,784
0,480	0,137	0,757	945,997
0,470	0,140	0,760	958,798
0,460	0,142	0,765	969,900
0,450	0,145	0,771	982,207
0,440	0,148	0,776	995,058
0,430	0,151	0,779	1010,778
0,420	0,154	0,786	1023,601
0,410	0,158	0,791	1039,207
0,400	0,161	0,797	1054,966
0,390	0,165	0,802	1073,839
0,380	0,169	0,807	1091,722
0,370	0,173	0,807	1118,575
0,360	0,178	0,816	1135,819
0,350	0,183	0,821	1163,983
0,340	0,189	0,826	1189,159
0,330	0,194	0,830	1218,804
0,320	0,201	0,836	1250,701
0,310	0,208	0,843	1283,843
0,300	0,216	0,853	1318,641
0,290	0,224	0,859	1361,784
0,280	0,234	0,867	1406,110
0,270	0,245	0,872	1464,853
0,260	0,259	0,878	1537,339
0,250	0,274	0,884	1616,513
0,240	0,289	0,891	1692,216
0,230	0,315	0,905	1815,451
0,220	0,338	0,913	1928,364
0,210	0,361	0,921	2043,782
0,200	0,396	0,938	2200,107
0,190	0,437	0,945	2408,020
0,180	0,478	0,954	2608,815
0,170	0,522	0,961	2830,011
0,160	0,562	0,971	3013,683
0,150	0,605	0,977	3225,893
0,140	0,661	0,985	3495,511
0,130	0,691	0,986	3649,232
0,120	0,713	0,990	3749,672
0,110	0,738	0,993	3874,560
0,100	0,765	0,998	3994,042
0,090	0,797	0,998	4159,968
0,080	0,833	0,999	4342,308
0,070	0,874	0,999	4558,461
0,060	0,927	1,000	4828,865
0,050	0,970	1,000	5052,098
0,040	0,985	1,000	5129,023
0,030	0,995	1,000	5182,218
0,020	0,999	1,000	5205,954
0,010	1,000	1,000	5208,642
0,000	1,000	1,000	5208,709

LASSO 2 años

Punto de corte (probabilidad de SI tener cáncer)	Porcentaje de población a intervenir	Porcentaje de pacientes con cáncer intervenidos	Número de personas a seguir
1,000	0,000	0,000	#¡DIV/0!
0,990	0,000	0,000	#¡DIV/0!
0,980	0,000	0,005	151,200
0,970	0,000	0,024	95,625
0,960	0,001	0,041	118,357
0,950	0,002	0,056	162,860
0,940	0,003	0,087	175,045
0,930	0,004	0,106	212,953
0,920	0,006	0,140	215,746
0,910	0,008	0,177	219,821
0,900	0,009	0,200	240,564
0,890	0,011	0,220	267,468
0,880	0,014	0,253	284,520
0,870	0,017	0,288	306,990
0,860	0,022	0,321	344,945
0,850	0,030	0,376	413,089
0,840	0,034	0,397	440,229
0,830	0,040	0,420	481,741
0,820	0,046	0,445	530,029
0,810	0,052	0,476	564,428
0,800	0,055	0,486	583,774
0,790	0,058	0,492	606,428
0,780	0,061	0,500	624,121
0,770	0,063	0,503	641,391
0,760	0,065	0,516	649,253
0,750	0,068	0,524	661,689
0,740	0,069	0,534	665,748
0,730	0,071	0,538	678,721
0,720	0,073	0,542	686,091
0,710	0,075	0,548	701,087
0,700	0,077	0,553	710,286
0,690	0,078	0,558	718,335
0,680	0,081	0,563	732,504
0,670	0,083	0,569	749,811
0,660	0,086	0,574	769,962
0,650	0,089	0,578	791,526
0,640	0,092	0,584	806,580
0,630	0,095	0,594	821,704
0,620	0,100	0,602	848,102
0,610	0,103	0,607	866,303
0,600	0,106	0,614	883,377
0,590	0,109	0,621	899,510
0,580	0,112	0,623	919,729
0,570	0,115	0,630	936,107
0,560	0,119	0,634	955,523
0,550	0,122	0,637	978,701
0,540	0,125	0,638	1001,073
0,530	0,128	0,646	1014,757
0,520	0,132	0,652	1031,742
0,510	0,135	0,662	1045,676
0,500	0,139	0,675	1053,442
0,490	0,143	0,682	1072,280
0,480	0,147	0,689	1092,928
0,470	0,152	0,693	1121,608
0,460	0,157	0,696	1155,651
0,450	0,163	0,707	1176,780
0,440	0,169	0,719	1202,120
0,430	0,175	0,728	1229,834
0,420	0,182	0,730	1271,963
0,410	0,190	0,741	1307,733
0,400	0,199	0,749	1360,078
0,390	0,211	0,764	1409,718
0,380	0,223	0,776	1472,222
0,370	0,240	0,794	1545,843
0,360	0,253	0,801	1612,088
0,350	0,270	0,813	1694,725
0,340	0,287	0,829	1768,219
0,330	0,314	0,848	1895,012
0,320	0,344	0,863	2034,502
0,310	0,376	0,876	2193,074
0,300	0,396	0,885	2288,561
0,290	0,420	0,896	2397,674
0,280	0,454	0,905	2566,633
0,270	0,482	0,912	2697,668
0,260	0,512	0,922	2835,547
0,250	0,549	0,936	2994,629
0,240	0,574	0,939	3122,723
0,230	0,597	0,943	3232,345
0,220	0,622	0,949	3349,332
0,210	0,646	0,957	3450,300
0,200	0,672	0,965	3560,797
0,190	0,697	0,968	3677,427
0,180	0,721	0,971	3789,260
0,170	0,741	0,973	3889,219
0,160	0,762	0,977	3984,626
0,150	0,782	0,981	4071,693
0,140	0,811	0,984	4211,329
0,130	0,835	0,987	4321,845
0,120	0,862	0,990	4445,222
0,110	0,888	0,990	4579,601
0,100	0,912	0,995	4683,543
0,090	0,936	0,997	4795,348
0,080	0,957	0,999	4891,506
0,070	0,973	0,999	4975,052
0,060	0,983	1,000	5019,456
0,050	0,990	1,000	5057,589
0,040	0,995	1,000	5083,922
0,030	0,998	1,000	5098,884
0,020	1,000	1,000	5106,686
0,010	1,000	1,000	5107,808
0,000	1,000	1,000	5107,825

Bosques aleatorios

Punto de corte (probabilidad de SI tener cáncer)	Porcentaje de población a intervenir	Porcentaje de pacientes con cáncer intervenidos	Número de personas a seguir
1,000	0,000	0,000	#¡DIV/0!
0,990	0,000	0,000	#¡DIV/0!
0,980	0,000	0,000	#¡DIV/0!
0,970	0,000	0,000	#¡DIV/0!
0,960	0,000	0,000	#¡DIV/0!
0,950	0,000	0,000	#¡DIV/0!
0,940	0,000	0,000	#¡DIV/0!
0,930	0,000	0,000	#¡DIV/0!
0,920	0,000	0,000	#¡DIV/0!
0,910	0,000	0,000	#¡DIV/0!
0,900	0,000	0,000	#¡DIV/0!
0,890	0,000	0,000	#¡DIV/0!
0,880	0,000	0,000	#¡DIV/0!
0,870	0,000	0,000	#¡DIV/0!
0,860	0,000	0,000	#¡DIV/0!
0,850	0,000	0,000	#¡DIV/0!
0,840	0,000	0,000	#¡DIV/0!
0,830	0,000	0,000	#¡DIV/0!
0,820	0,000	0,000	#¡DIV/0!
0,810	0,000	0,000	#¡DIV/0!
0,800	0,000	0,000	#¡DIV/0!
0,790	0,000	0,000	#¡DIV/0!
0,780	0,000	0,000	#¡DIV/0!
0,770	0,000	0,000	#¡DIV/0!
0,760	0,000	0,000	#¡DIV/0!
0,750	0,000	0,000	#¡DIV/0!
0,740	0,000	0,000	#¡DIV/0!
0,730	0,000	0,001	63,000
0,720	0,000	0,003	87,500
0,710	0,000	0,008	68,417
0,700	0,000	0,017	72,040
0,690	0,000	0,027	85,463
0,680	0,001	0,037	108,696
0,670	0,001	0,052	128,167
0,660	0,002	0,071	145,991
0,650	0,003	0,093	167,186
0,640	0,004	0,111	201,732
0,630	0,006	0,140	225,019
0,620	0,008	0,166	259,820
0,610	0,011	0,206	277,574
0,600	0,015	0,250	298,524
0,590	0,019	0,289	334,421
0,580	0,024	0,332	372,611
0,570	0,030	0,374	417,511
0,560	0,039	0,425	467,670
0,550	0,049	0,474	533,254
0,540	0,062	0,526	607,210
0,530	0,077	0,569	699,163
0,520	0,094	0,605	801,362
0,510	0,114	0,641	915,996
0,500	0,139	0,675	1055,154
0,490	0,169	0,706	1230,975
0,480	0,205	0,732	1438,636
0,470	0,249	0,762	1676,704
0,460	0,303	0,802	1938,104
0,450	0,365	0,838	2238,933
0,440	0,434	0,864	2580,575
0,430	0,511	0,898	2926,874
0,420	0,597	0,923	3320,548
0,410	0,681	0,942	3713,353
0,400	0,759	0,958	4065,376
0,390	0,826	0,975	4352,564
0,380	0,885	0,987	4600,972
0,370	0,929	0,994	4796,398
0,360	0,962	0,997	4951,830
0,350	0,985	0,998	5067,306
0,340	0,997	0,999	5119,743
0,330	1,000	1,000	5130,861
0,320	1,000	1,000	5132,093
0,310	1,000	1,000	5132,262
0,300	1,000	1,000	5132,270
0,290	1,000	1,000	5132,270
0,280	1,000	1,000	5132,270
0,270	1,000	1,000	5132,270
0,260	1,000	1,000	5132,270
0,250	1,000	1,000	5132,270
0,240	1,000	1,000	5132,270
0,230	1,000	1,000	5132,270
0,220	1,000	1,000	5132,270
0,210	1,000	1,000	5132,270
0,200	1,000	1,000	5132,270
0,190	1,000	1,000	5132,270
0,180	1,000	1,000	5132,270
0,170	1,000	1,000	5132,270
0,160	1,000	1,000	5132,270
0,150	1,000	1,000	5132,270
0,140	1,000	1,000	5132,270
0,130	1,000	1,000	5132,270
0,120	1,000	1,000	5132,270
0,110	1,000	1,000	5132,270
0,100	1,000	1,000	5132,270
0,090	1,000	1,000	5132,270
0,080	1,000	1,000	5132,270
0,070	1,000	1,000	5132,270
0,060	1,000	1,000	5132,270
0,050	1,000	1,000	5132,270
0,040	1,000	1,000	5132,270
0,030	1,000	1,000	5132,270
0,020	1,000	1,000	5132,270
0,010	1,000	1,000	5132,270
0,000	1,000	1,000	5132,270

Potenciación del gradiente

Punto de corte (probabilidad de SI tener cáncer)	Porcentaje de población a intervenir	Porcentaje de pacientes con cáncer intervenidos	Número de personas a seguir
1,000	0,000	0,000	#¡DIV/0!
0,990	0,000	0,000	#¡DIV/0!
0,980	0,000	0,000	#¡DIV/0!
0,970	0,000	0,000	#¡DIV/0!
0,960	0,000	0,000	#¡DIV/0!
0,950	0,000	0,000	#¡DIV/0!
0,940	0,000	0,001	129,500
0,930	0,000	0,008	105,250
0,920	0,001	0,030	128,043
0,910	0,002	0,056	150,671
0,900	0,004	0,103	197,019
0,890	0,008	0,151	270,035
0,880	0,011	0,189	296,270
0,870	0,014	0,224	322,282
0,860	0,019	0,261	371,612
0,850	0,024	0,293	418,176
0,840	0,033	0,347	492,615
0,830	0,045	0,400	577,226
0,820	0,052	0,425	625,956
0,810	0,055	0,449	632,536
0,800	0,059	0,462	655,262
0,790	0,061	0,472	664,867
0,780	0,063	0,479	677,169
0,770	0,066	0,488	694,054
0,760	0,068	0,495	701,469
0,750	0,069	0,502	710,416
0,740	0,072	0,508	725,758
0,730	0,073	0,514	732,304
0,720	0,076	0,517	750,316
0,710	0,078	0,523	762,295
0,700	0,079	0,527	773,501
0,690	0,081	0,532	782,923
0,680	0,083	0,537	797,402
0,670	0,085	0,544	804,337
0,660	0,087	0,548	815,229
0,650	0,089	0,549	828,257
0,640	0,090	0,552	840,553
0,630	0,093	0,558	853,068
0,620	0,095	0,564	861,851
0,610	0,098	0,573	882,446
0,600	0,101	0,577	896,860
0,590	0,103	0,579	913,695
0,580	0,106	0,584	934,985
0,570	0,110	0,594	947,187
0,560	0,112	0,600	959,857
0,550	0,115	0,605	976,467
0,540	0,120	0,611	1003,955
0,530	0,124	0,620	1024,453
0,520	0,128	0,626	1048,863
0,510	0,131	0,630	1069,050
0,500	0,134	0,636	1084,879
0,490	0,138	0,642	1104,274
0,480	0,145	0,650	1141,956
0,470	0,151	0,658	1181,301
0,460	0,157	0,668	1209,274
0,450	0,162	0,671	1240,796
0,440	0,167	0,678	1266,219
0,430	0,175	0,684	1314,841
0,420	0,182	0,690	1357,237
0,410	0,189	0,695	1391,972
0,400	0,196	0,699	1439,277
0,390	0,205	0,701	1498,048
0,380	0,211	0,705	1538,209
0,370	0,223	0,717	1601,163
0,360	0,233	0,724	1649,411
0,350	0,247	0,736	1724,290
0,340	0,261	0,745	1798,291
0,330	0,284	0,759	1924,887
0,320	0,309	0,776	2044,199
0,310	0,327	0,786	2133,795
0,300	0,346	0,792	2244,174
0,290	0,366	0,805	2333,967
0,280	0,391	0,819	2454,434
0,270	0,408	0,827	2536,765
0,260	0,421	0,834	2595,941
0,250	0,444	0,841	2710,767
0,240	0,461	0,846	2800,712
0,230	0,478	0,857	2862,444
0,220	0,502	0,866	2975,476
0,210	0,538	0,876	3151,691
0,200	0,568	0,888	3282,080
0,190	0,601	0,899	3431,909
0,180	0,626	0,907	3541,466
0,170	0,665	0,919	3710,991
0,160	0,692	0,927	3832,248
0,150	0,723	0,939	3950,075
0,140	0,747	0,952	4031,368
0,130	0,773	0,963	4121,775
0,120	0,798	0,971	4219,279
0,110	0,830	0,979	4352,414
0,100	0,878	0,981	4589,443
0,090	0,914	0,989	4744,286
0,080	0,962	0,994	4966,261
0,070	0,984	0,997	5067,355
0,060	0,996	0,997	5128,079
0,050	0,999	0,999	5135,208
0,040	1,000	1,000	5131,780
0,030	1,000	1,000	5132,624
0,020	1,000	1,000	5132,780
0,010	1,000	1,000	5132,806
0,000	1,000	1,000	5132,807

Anexo 6. Análisis exploratorio en cáncer de páncreas

El procesamiento de datos fue apoyado por Lizeth Roldán como componente de su trabajo de grado en la Maestría en Ingeniería Electrónica de la Pontificia Universidad Javeriana.

Se utilizaron los registros del año 2013; el año se seleccionó con base en el análisis de calidad de los datos descrito en los métodos. La preparación de los datos siguió los mismos pasos descritos para el análisis realizado con cáncer de estómago. Se revisó el etiquetado de cáncer páncreas solo mediante análisis exploratorio, no se realizó validación con registros clínicos ni con el Registro Poblacional de Cáncer de Bucaramanga.

Igual que en cáncer gástrico, para el análisis exploratorio del etiquetado se usaron 5 categorías que combinan diagnóstico de cáncer de páncreas (CIE10), enfermedad benigna del páncreas con procedimientos de cáncer de páncreas (CIE10-CUPS), procedimientos de cáncer en general con enfermedad benigna del páncreas, medicamentos de cáncer de páncreas con enfermedad benigna del páncreas, y medicamentos oncológicos en general con enfermedad benigna de páncreas.

La codificación utilizada se describe a continuación:

1) Código CIE10 cáncer de páncreas
    C240 Tumor maligno de las vías biliares extrahepáticas
    C241 Tumor maligno de la ampolla de váter
    C250 Tumor maligno de la cabeza del páncreas
    C251 Tumor maligno del cuerpo del páncreas
    C252 Tumor maligno de la cola del páncreas
    C253 Tumor maligno del conducto pancreático
    C254 Tumor maligno del páncreas endocrino
    C257 Tumor maligno de otras partes especificadas del páncreas
    C258 Lesión de sitios contiguos del páncreas
    C259 Tumor maligno del páncreas, parte no especificada
    Z85.07 Historia personal de neoplasia maligna del páncreas
2) Código CIE10 enfermedad benigna del páncreas
    K85X Pancreatitis aguda
    K85 Pancreatitis aguda
    K850 Pancreatitis idiopática aguda
    K851 Pancreatitis biliar aguda
    K852 Pancreatitis aguda inducida por alcohol
    K853 Pancreatitis aguda inducida por drogas
    K858 Otras pancreatitis agudas
    K859 Pancreatitis aguda, no especificada
    K86 Otras enfermedades del páncreas
    K860 Pancreatitis crónica inducida por el alcohol
    K861 Otras pancreatitis crónicas
    K862 Quiste del páncreas

K863 Pseudoquiste del páncreas
K868 Otras enfermedades especificadas del páncreas
K869 Enfermedad del páncreas, no especificada

3) Código CIE10 cáncer general:

Todos los códigos del CIE-10 de cáncer excluyendo cáncer de páncreas

4) Procedimientos relacionados con cáncer de páncreas (CUPS)

52.5.0 Pancreatectomía central
52.5 Pancreatectomía parcial
52.6 Pancreatectomía total
52.6.1.00 Pancreatectomía total por necidioblastosis
52.5.0.01 Pancreatectomía central vía abierta
51.1.0 Pancreatectomía total vía abierta
52.5.0.02 Pancreatectomía central vía laparoscópica
51.1.0.00 Pancreatectomía total vía laparoscópica
52.5.1 Pancreatectomía proximal
52.1.2 Pancreatectomía total por necidioblastosis SOD
52.5.1.00 Pancreatectomía proximal SOD
52.1.2 Colangiopancratografía retrógrada endoscópica CPRE
52.5.1.00 Pancreatectomía proximal vía abierta
52.1.2 Colangiopancratografía retrógrada endoscópica CPRE SOD
52.5.1.00 Pancreatectomía proximal vía laparoscópica
52.1.2 Biopsias de páncreas
52.5.2 Pancreatectomía distal
52.1.2 Biopsias de páncreas vía abierta
52.5.2.00 Pancreatectomía distal SOD
52.1.1 Biopsias de páncreas vía percutánea
52.5.2 Pancreatectomía distal vía abierta
52.1.1.00 Biopsias de páncreas vía laparoscópica
52.5.2.2.04 Pancreatectomía distal vía laparoscópica
52.1.2 Biopsias de páncreas vía endoscópica
52.5.2.2.05 Pancreatectomía distal con preservación de bazo vía abierta
52.1.2.00 Biopsia por aspiración (aguja) cerrada de páncreas
52.5.3 Pancreatectomía subtotal (operación de CHILD)
52.1.3.00 Biopsia abierta de páncreas
52.5.3.01 Pancreatectomía subtotal vía abierta
52.6 Biopsia abierta de páncreas SOD
52.5.3.02 Pancreatectomía subtotal vía laparoscópica
52.6.1.00 Pancreatografía retrógrada endoscópica ERP
52.5.3.00 Pancreatectomía subtotal (operación de CHILD) SOD
51.1.0 Pancreatografía retrógrada endoscópica ERP SOD
89.0.2.34 Consulta de primera vez por especialista en cirugía gastrointestinal

5) Medicamentos relacionados con cáncer de páncreas (código ATC)

L01BC06 / Capecitabina
L01BC05 / Gemcitabina
L01XA01 / Cisplatino
L01XX19 / Irinotecan

L01XA03 / Oxaliplatino
L01CD0101 / Paclitaxel
L01XE03 / Erlotinib
L01XE53 / Larotrectinib
L01XC18 / Pembrolizumab

Con base en los resultados del etiquetado se seleccionaron casos (cáncer de páncreas) y controles (sin cáncer de páncreas) de los pacientes atendidos en diciembre de 2013.

Se realizó un ejercicio de minería de patrones (conglomerados de procesos) para ver rutas clínicas relacionadas con el cáncer de páncreas, esto incluye descubrimiento de procesos, verificación de conformidad de acuerdo con el proceso descubierto y/o modelado, y mejora del proceso; sin embargo, este último componente no fue desarrollado ya que no corresponde al objetivo del análisis. El modelo se basa en la reconstrucción del historial clínico para casos y controles, lo que permite asociar diagnósticos y temporalidad del diagnóstico a fin de reconstruir la ruta asistencial de cada paciente, para luego agruparla y generar procesos que representen a todos los pacientes en el conglomerado. El historial clínico se reconstruyó con el historial de diagnósticos de enero a noviembre de 2013.

Adicionalmente, para identificar variables con potencial predictivo se utilizó regresión logística LASSO, de igual forma a lo realizado en el modelo de cáncer gástrico. Se analizaron coeficientes en un proceso de validación cruzada de 10 grupos, repetido 10 veces.

Las variables incluidas en el análisis son el identificador, fecha de atención, fecha de nacimiento (edad al diagnóstico), aseguradora (EPS), diagnóstico (CIE10), sexo, procedimientos (CUPS), municipio (DIVIPOLA), medicamentos (CUMS-ATC), centro de atención (Código de Habilitación). Cada código al interior de estas clasificaciones se transformó en una variable binaria. Se eliminaron los registros sin información para las variables descritas.

Se filtró la información por revisión manual de los códigos asociados buscando plausibilidad teórica de los hallazgos, a fin de reducir el número de variables y el costo computacional asociado. Se eliminaron diagnósticos relacionados con cáncer en el periodo enero a noviembre de 2013. Adicionalmente, se eliminaron los siguientes diagnósticos relacionados con síndromes para-neoplásicos, procedimientos oncológicos o efectos adversos de tratamiento del cáncer:

G131-Otras atrofias sistémicas que afectan el sistema nervioso central en enfermedad neoplásica
G533-Parálisis múltiple de los nervios craneales
G631-Polineuropatia en enfermedad neoplásica
G732-Otros síndromes miasténicos en enfermedad neoplásica
T860-Rechazo de trasplante de medula ósea
Y431-Efectos adversos de antimetabolitos antineoplásicos

Y433-Efectos adversos de otras drogas antineoplásicas
Z091-Examen de seguimiento consecutivo a radioterapia por otras afecciones
Z092-Examen de seguimiento consecutivo a quimioterapia por otras afecciones
Z510-Sesión de radioterapia
Z511-Sesión de quimioterapia por tumor
Z512-Otra quimioterapia
Z515-Atención paliativa
Z541-Convalecencia consecutiva a radioterapia
Z542-Convalecencia consecutiva a quimioterapia
Z801-Historia familiar de tumor maligno de tráquea
Z807-Historia familiar de otros tumores malignos del tejido linfoide
Z808-Historia familiar de tumor maligno de otros órganos o sistemas especificados
Z851-Historia personal de tumor maligno de tráquea
Z852-Historia personal de tumor maligno de otros órganos respiratorios e intra-torácicos
Z855-Historia personal de tumor maligno de vías urinarias
Z857-Historia personal de otros tumores malignos del tejido linfoide
Z858-Historia personal de tumor maligno de otros órganos y sistemas

El mismo procedimiento se siguió para los medicamentos y procedimientos.

Se utilizaron diferentes técnicas de aprendizaje de máquinas para la construcción de modelos predictivos (aprendizaje supervisado): LASSO, bosques aleatorios (BA), redes neuronales artificiales (RNA), y máquina de soporte vectorial (MSV); sin embargo, en este último no se logró su convergencia. Cada algoritmo se entrenó con partición de datos de 70%-30% para entrenamiento y prueba respectivamente.

Los hiper-parámetros más comunes se ajustaron mediante sintonización con 2000 muestras aleatorias. Para el entrenamiento de los algoritmos de BA, MSV y RNA se utilizaron las variables con coeficiente diferente de cero reportadas por la regresión logística LASSO.

Tabla 6-1. Parámetros para la sintonización de modelos predictivos

Modelo	Parámetro	Valor
Regresión LASSO	Constante de penaización	0,01 a 0,09 (paso = 0,01)
Bosques aleatorios	Función de costo	*Gini o entropy*
	Número de árboles	50 a 100 (paso = 10)
	Características del nodo	*auto, sqrt y log 2*
	Máxima profundidad	10 a 110 (paso = 10)
	Mínimo de muestras	0,1 a 1 (paso = 0,1)
Máquina de soporte vectorial	Kernels	*Kernel Gaussiano, Kernel Sigmoide o Kernel Polinomial*
	Regularización	1 a 100 (paso = 1)
	Gamma	*Gamma alta y gamma baja*
Redes neuronales artificiales	Función de activación	*relu, tanh, sigmoid*
	Cantidad de neuronas por capa	5 a 145 (múltiplos de 5)

Resultados

Para el año 2013 se identificaron en total 5.451.496 pacientes, 9.638 diagnósticos, 4.840 695 procedimientos y 695 medicamentos.

El análisis exploratorio del etiquetado mostró poco valor del uso de medicamentos para este fin

Tabla 6-2. Resultados análisis exploratorio de etiquetado cáncer de páncreas

Etiquetado	No. registros hallados
CIE10 cáncer de páncreas	20399
Procedimientos cáncer de páncreas y enfermedad benigna del páncreas	1754
Procedimientos cáncer en general y enfermedad benigna del páncreas	1512
Medicamentos cáncer de páncreas y enfermedad benigna del páncreas	0
Medicamentos cáncer en general y enfermedad benigna del páncreas	8

Con base en los resultados observados, para los modelos predictivos se decide utilizar la combinación de CIE10 cáncer de páncreas y CUPS de procedimientos cáncer de páncreas, excluyendo los códigos de enfermedad benigna.

En el análisis de conglomerados se observaron procesos que agrupan la mayoría de rutas asistenciales para pacientes con cáncer de páncreas comparado con pacientes sin cáncer. Se encontraron 25 conglomerados; sin embargo, el aporte de datos agrupados no permite identificar de forma discriminada la capacidad predictiva de las variables, por ello se utiliza LASSO para este propósito la cual fue fácilmente interpretable. Las variables con valor predictivo identificadas en el análisis preliminar se pueden categorizar como sigue:

1. Diagnósticos:
    a. Diagnósticos diferenciales de cáncer de páncreas (ej. tumor benigno, cálculo en vía biliar, etc.)
    b. Expresión sintomática de la enfermedad (ej. dolor abdominal, dispepsia, etc.)
    c. Condiciones asociadas a la enfermedad (ej. diabetes mellitus, celulitis, anemia, etc.)
    d. Condiciones asociadas a factores de riesgo para la enfermedad (ej. infección de vías respiratorias)
    e. Otros no relacionables

2. Medicamentos más significativos
    a. Antibióticos
    b. Medicamentos para trastornos hematológicos
    c. Manejo de síntomas intestinales
    d. Manejo de síndrome metabólico
    e. Anti-arrítmicos

3. Procedimientos significativos
    a. Diagnóstico (ej. Resonancia nuclear magnética de abdomen, ecografía abdominal, biopsia, etc.)
    b. Extensión y condición biológica (ej. endoscopia, hemograma, etc.)

En total se realizaron 45 iteraciones de la regresión logística LASSO. Las mejores áreas bajo la curva ROC (AUCROC) para este modelo se encontraron con constantes de penalización entre 0,01 y 0,08

C=0.01 con un AUC de 0.71
C=0.02 con un AUC de 0.63
C=0.04 con un AUC de 0.68
C=0.08 con un AUC de 0.64

En el modelo de bosques aleatorios, una vez finalizada la sintonización del modelo se obtuvieron los siguientes valores para cada uno de los parámetros:

Función de costo: Gini
Número de estimadores: 800
Características del nodo: sqrt
Máxima profundidad del árbol: 90
Muestras mínimas del split: 0.44
Máxima profundidad del árbol: 0.09

La AUCROC del modelo con mejor rendimiento en bosques aleatorios fue 0,72.

El modelo de redes neuronales se sintonizó variando la cantidad de neuronas, la tasa de aprendizaje y el número de capas ocultas. Los resultados obtenidos son

Capas ocultas 0 a 3
Cantidad de neuronas 1 a 700
Tasa de aprendizaje 0.01 a 0.03

La AUROC para este modelo fue 0,50 debido a que no logra capturar ningún paciente etiqueta de cáncer de páncreas.

Conclusiones

El desempeño de los modelos predictivos en cáncer de páncreas muestra diferencias frente al desempeño en la predicción de cáncer gástrico. El bajo desempeño de las redes neuronales y la no convergencia del modelo de máquinas de soporte vectorial sugiere el uso de regresión logística LASSO y bosques aleatorios como modelos prioritarios en la construcción de modelos predictivos con las utilizando las bases de datos del estudio.

Sería conveniente a futuro revisar la construcción del modelo siguiendo más fielmente la metodología utilizada para la modelación de cáncer gástrico incluyendo un periodo de observación más prolongado y el periodo de lavado para el análisis del último año previo al diagnóstico.

13. Bibliografía

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229-63.
2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2018;68(6):394-424.
3. Fondo Colombiano de Enfermedades de Alto Costo. Situación del cáncer en la población adulta atendida en el SGSSS en Colombia. 2022. Bogotá DC, Cuenta de Alto Costo (CAC); 2023.
4. Allemani C, Matsuda T, Di Carlo V, Harewood R, Matz M, Niksic M, et al. Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. Lancet. 2018;391(10125):1023-75.
5. Langton JM, Blanch B, Drew AK, Haas M, Ingham JM, Pearson SA. Retrospective studies of end-of-life resource utilization and costs in cancer care using health administrative data: a systematic review. Palliat Med. 2014;28(10):1167-96.
6. Bravo LE, Hernández Vargas JA, Collazos P, García LS, Valbuena AM, Acuña L. Survival in stomach cancer: analysis of a national cancer information system and a population-based cancer registry in Colombia. Colombia Médica. 2022;53(4).
7. Zhang X, Li M, Chen S, Hu J, Guo Q, Liu R, et al. Endoscopic Screening in Asian Countries Is Associated With Reduced Gastric Cancer Mortality: A Meta-analysis and Systematic Review. Gastroenterology. 2018;155(2):347-54 e9.
8. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2021;71(3):209-49.
9. Murillo R. Cancer screening: from biology to public health. World Cancer Report: Cancer Research for Cancer Prevention. Lyon, France: International Agency for Research on Cancer; 2020. p. 540-9.
10. Malfertheiner P, Camargo MC, El-Omar E, Liou JM, Peek R, Schulz C, et al. Helicobacter pylori infection. Nat Rev Dis Primers. 2023;9(1):19.
11. IARC Helicobacter pylori Working Group 2014. Helicobacter pylori Eradication as a Strategy for Preventing Gastric Cancer. Lyon: International Agency for research on Cancer; 2014 Dec 2013.
12. Charvat H, Sasazuki S, Inoue M, Iwasaki M, Sawada N, Shimazu T, et al. Prediction of the 10-year probability of gastric cancer occurrence in the Japanese population: the JPHC study cohort II. Int J Cancer. 2016;138(2):320-31.
13. Eom BW, Joo J, Kim S, Shin A, Yang HR, Park J, et al. Prediction Model for Gastric Cancer Incidence in Korean Population. PLoS One. 2015;10(7):e0132613.
14. Hamilton W. The CAPER studies: five case-control studies aimed at identifying and quantifying the risk of cancer in symptomatic primary care patients. British journal of cancer. 2009;101(2):S80-S6.
15. Niu PH, Zhao LL, Wu HL, Zhao DB, Chen YT. Artificial intelligence in gastric cancer: Application and future perspectives. World J Gastroenterol. 2020;26(36):5408-19.
16. Cao R, Tang L, Fang M, Zhong L, Wang S, Gong L, et al. Artificial intelligence in gastric cancer: applications and challenges. Gastroenterol Rep (Oxf). 2022;10:goac064.
17. Fan Z, He Z, Miao W, Huang R. Critical Analysis of Risk Factors and Machine-Learning-Based Gastric Cancer Risk Prediction Models: A Systematic Review. Processes. 2023;11(8):2324.

18. Afrash MR, Shafiee M, Kazemi-Arpanahi H. Establishing machine learning models to predict the early risk of gastric cancer based on lifestyle factors. BMC Gastroenterol. 2023;23(1):6.
19. Briggs E, de Kamps M, Hamilton W, Johnson O, McInerney CD, Neal RD. Machine Learning for Risk Prediction of Oesophago-Gastric Cancer in Primary Care: Comparison with Existing Risk-Assessment Tools. Cancers (Basel). 2022;14(20).
20. Mohammadnezhad K, Sahebi MR, Alatab S, Sadjadi A. Modeling Epidemiology Data with Machine Learning Technique to Detect Risk Factors for Gastric Cancer. J Gastrointest Cancer. 2023.
21. Huang RJ, Kwon NS, Tomizawa Y, Choi AY, Hernandez-Boussard T, Hwang JH. A Comparison of Logistic Regression Against Machine Learning Algorithms for Gastric Cancer Risk Prediction Within Real-World Clinical Data Streams. JCO Clin Cancer Inform. 2022;6:e2200039.
22. International Agency for Reseach on Cancer-World Health Organization. Global Cancer Observatory 2022. Available from: http://gco.iarc.fr.
23. Rose G. Sick individuals and sick populations. International journal of epidemiology. 2001;30(3):427-32.
24. Hu B, El Hajj N, Sittler S, Lammert N, Barnes R, Meloni-Ehrig A. Gastric cancer: Classification, histology and application of molecular pathology. Journal of gastrointestinal oncology. 2012;3(3):251.
25. Sierra MS, Cueva P, Bravo LE, Forman D. Stomach cancer burden in Central and South America. Cancer epidemiology. 2016;44:S62-S73.
26. Malnick SDH, Melzer E, Attali M, Duek G, Yahav J. Helicobacter pylori: friend or foe? World journal of gastroenterology: WJG. 2014;20(27):8979.
27. Bravo LE, Cortés A, Carrascal E, Jaramillo R, García LS, Bravo PE, et al. Helicobacter pylori: patología y prevalencia en biopsias gástricas en Colombia. Colombia médica. 2003;34(3):124-31.
28. Kandulski A, Malfertheiner P, Wex T. Role of regulatory T-cells in H. pylori-induced gastritis and gastric cancer. Anticancer research. 2010;30(4):1093-103.
29. Kayali S, Manfredi M, Gaiani F, Bianchi L, Bizzarri B, Leandro G, et al. Helicobacter pylori, transmission routes and recurrence of infection: state of the art. Acta Biomed. 2018;89(8-S):72-6.
30. Arévalo A, Trespalacios AA, Otero W. The importance of CagA protein in Helicobacter Pylori infection. Revista colombiana de Gastroenterología. 2009;24(4):388-95.
31. Tempera PJ, Michael M, Tageldin O, Hasak S. Gastric cancer due to chronic H. pylori infection: what we know and where we are going. Diseases. 2022;10(3):57.
32. Dincă AL, Meliț LE, Mărginean CO. Old and new aspects of H. pylori-associated Inflammation and gastric cancer. Children. 2022;9(7):1083.
33. Fielding J, Ellis D, Jones B, Paterson J, Powell D, Waterhouse J, et al. Natural history of" early" gastric cancer: results of a 10-year regional survey. Br Med J. 1980;281(6246):965-7.
34. Tsukuma H, Oshima A, Narahara H, Morii T. Natural history of early gastric cancer: a non-concurrent, long term, follow up study. Gut. 2000;47(5):618-21.
35. Nadler DL, Zurbenko IG. Estimating cancer latency times using a Weibull model. Advances in Epidemiology. 2014;2014.
36. Iida T, Yamashita K, Ohwada S, Ohkubo Y, Hirano T, Miyake T, et al. Natural history of gastric cancer from a retrospective review of endoscopic images of older patients with interval gastric cancer. Geriatrics & Gerontology International. 2018;18(7):997-1002.
37. Huang RJ, Laszkowska M, In H, Hwang JH, Epplein M. Controlling Gastric Cancer in a World of Heterogeneous Risk. Gastroenterology. 2023;164(5):736-51.
38. Januszewicz W, Turkot MH, Malfertheiner P, Regula J. A Global Perspective on Gastric Cancer Screening: Which Concepts Are Feasible, and When? Cancers. 2023;15(3):664.

39. Hibino M, Hamashima C, Iwata M, Terasawa T. Radiographic and endoscopic screening to reduce gastric cancer mortality: a systematic review and meta-analysis. The Lancet Regional Health–Western Pacific. 2023.
40. Terasawa T, Nishida H, Kato K, Miyashiro I, Yoshikawa T, Takaku R, et al. Prediction of gastric cancer development by serum pepsinogen test and Helicobacter pylori seropositivity in Eastern Asians: a systematic review and meta-analysis. PloS one. 2014;9(10):e109783.
41. Zagari RM, Rabitti S, Greenwood DC, Eusebi LH, Vestito A, Bazzoli F. Systematic review with meta-analysis: diagnostic performance of the combination of pepsinogen, gastrin-17 and anti-Helicobacter pylori antibodies serum assays for the diagnosis of atrophic gastritis. Alimentary pharmacology & therapeutics. 2017;46(7):657-67.
42. Robles C, Rudzite D, Polaka I, Sjomina O, Tzivian L, Kikuste I, et al. Assessment of serum pepsinogens with and without co-testing with gastrin-17 in gastric cancer risk assessment—Results from the GISTAR pilot study. Diagnostics. 2022;12(7):1746.
43. Necula L, Matei L, Dragu D, Neagu AI, Mambet C, Nedeianu S, et al. Recent advances in gastric cancer early diagnosis. World journal of gastroenterology. 2019;25(17):2029.
44. Calanzani N, Druce PE, Snudden C, Milley KM, Boscott R, Behiyat D, et al. Identifying Novel Biomarkers Ready for Evaluation in Low-Prevalence Populations for the Early Detection of Upper Gastrointestinal Cancers: A Systematic Review. Adv Ther. 2021;38(2):793-834.
45. Oliveros-Wilches R, Grillo-Ardila CF, Vallejo-Ortega M, Gil-Parada F, Cardona-Tobón M, Páramo-Hernández D, et al. Guía de práctica clínica para la prevención primaria y secundaria y diagnóstico temprano de cáncer gástrico. Revista Colombiana de Cancerología. 2022;26(1):39-96.
46. Gu J, Chen R, Wang S-M, Li M, Fan Z, Li X, et al. Prediction models for gastric cancer risk in the general population: a systematic review. Cancer Prevention Research. 2022;15(5):309-18.
47. Taninaga J, Nishiyama Y, Fujibayashi K, Gunji T, Sasabe N, Iijima K, et al. Prediction of future gastric cancer risk using a machine learning algorithm and comprehensive medical check-up data: A case-control study. Scientific reports. 2019;9(1):12384.
48. Leung WK, Cheung KS, Li B, Law SY, Lui TK. Applications of machine learning models in the prediction of gastric cancer risk in patients after Helicobacter pylori eradication. Alimentary pharmacology & therapeutics. 2021;53(8):864-72.
49. Hamilton W, Walter FM, Rubin G, Neal RD. Improving early diagnosis of symptomatic cancer. Nature Reviews Clinical Oncology. 2016;13(12):740-9.
50. Hippisley-Cox J, Coupland C. Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study. BMJ open. 2015;5(3):e007825.
51. Forster AS, Renzi C, Lyratzopoulos G. Diagnosing cancer in patients with 'non-alarm'symptoms: Learning from diagnostic care innovations in Denmark. Cancer epidemiology. 2018;54:101.
52. Van Es N, Le Gal G, Otten H-M, Robin P, Piccioli A, Lecumberri R, et al. Screening for occult cancer in patients with unprovoked venous thromboembolism: a systematic review and meta-analysis of individual patient data. Annals of internal medicine. 2017;167(6):410-7.
53. Koo MM, Hamilton W, Walter FM, Rubin GP, Lyratzopoulos G. Symptom signatures and diagnostic timeliness in cancer patients: a review of current evidence. Neoplasia. 2018;20(2):165-74.
54. Jutte DP, Roos LL, Brownell MD. Administrative record linkage as a tool for public health research. Annual review of public health. 2011;32:91-108.
55. Smith M, Lix LM, Azimaee M, Enns JE, Orr J, Hong S, et al. Assessing the quality of administrative data for research: a framework from the Manitoba Centre for Health Policy. Journal of the American Medical Informatics Association. 2018;25(3):224-9.

56. Benchimol EI. Epidemiology and health administrative data: Focus on methodology and transparency. Oxford University Press Oxford, UK; 2014. p. 1780-1.
57. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio) marker. Heart. 2012;98(9):683-90.
58. Ranstam J, Cook J, Collins G. Clinical prediction models. Journal of British Surgery. 2016;103(13):1886.
59. Lee Y-h, Bang H, Kim DJ. How to establish clinical prediction models. Endocrinology and Metabolism. 2016;31(1):38-44.
60. Shipe ME, Deppen SA, Farjah F, Grogan EL. Developing prediction models for clinical use using logistic regression: an overview. Journal of thoracic disease. 2019;11(Suppl 4):S574.
61. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691-8.
62. Vickers AJ. Prediction models in cancer care. CA: a cancer journal for clinicians. 2011;61(5):315-26.
63. Usher-Smith J, Emery J, Hamilton W, Griffin SJ, Walter FM. Risk prediction tools for cancer in primary care. British journal of cancer. 2015;113(12):1645-50.
64. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal. 2015;13:8-17.
65. Géron A. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: "O'Reilly Media, Inc."; 2022.
66. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology. 1996;58(1):267-88.
67. Ramos Castillo L. Regresión Lasso. 2018. Disponible en: https://idus.us.es/items/5a8cceef-1a26-4f46-b38c-009430551db6
68. Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review. 2021;54:1937-67.
69. Martínez Celda B. Gradient boosting (Potenciación del gradiente) en aprendizaje estadístico. 2021. Disponible en: https://uvadoc.uva.es/handle/10324/50607
70. Agrawal S, Agrawal J. Neural network techniques for cancer prediction: A survey. Procedia Computer Science. 2015;60:769-74.
71. Shell J, Gregory WD. Efficient cancer detection using multiple neural networks. IEEE journal of translational engineering in health and medicine. 2017;5:1-7.
72. Steiner MC. Solving cancer: The use of artificial neural networks in cancer diagnosis and treatment. Journal of Young Investigators. 2017;33(5):122-4.
73. Universidad Autónoma de Bucaramanga. Registro poblacional de cancer del área metropolitana de Bucaramanga. Vigencia 2021 2021 Disponible en: https://apolo.unab.edu.co/es/projects/registro-poblacional-de-cancer-del-area-metropolitana-de-bucarama.
74. Guerrero R, Gallego AI, Becerril-Montekio V, Vásquez J. Sistema de salud de Colombia. Salud pública de México. 2011;53:s144-s55.
75. Ministerio de Salud y Protección Social. Serie Anual de Afiliación a Nivel Nacional 1995-2022. In: Subdirección de la Operación del Aseguramiento en Salud, editor. República de Colombia2023.
76. Ministerio de Salud y Protección Social. Solicitud de información a Entidades Promotoras de Salud para el Estudio de Suficiencia y de los mecanismos de ajuste de riesgo para el cálculo de la

UPC recursos para garantizar la financiación de los servicios y tecnologías en salud año 2022. Periodo de estudio 2020. Bogotá, Colombia: Minsalud; 2020.
77. Ministerio de Salud y ¨Protección Social. Resolución 1896 de 2001: Por la cual se sustituye el anexo técnico de la Resolución 02333 de 2000, 1896 (2001). Bogotá, Colombia: Minsalud; 2001
78. Ministerio de Salud y Protección Social. Resolución 0255 DE 2007: Por la cual se adopta el Código Único Nacional de Medicamentos, CUM, 0255 (2007). Bogotá, Colombia: Minsalud; 2007
79. World Health Organization. ATC/DDD classification (temporary). WHO Drug Information. 2019;33(2):226-32.
80. Ministerio de Salud y Protección Social. Portal SISPRO: RUAFND – Módulo de nacimientos y defunciones. Minsalud; Disponible en: https://www.sispro.gov.co/central-prestadores-de-servicios/Pages/RUAFND.aspx.
81. Ajani JA, Bentrem DJ, Besh S, D'Amico TA, Das P, Denlinger C, et al. Gastric cancer, version 2.2013. Journal of the national comprehensive cancer network. 2013;11(5):531-46.
82. Winkler WE. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. U.S. Bureau of the Census - Stat. Research Div. Washington DC; 1990.
83. Damerau FJ. A technique for computer detection and correction of spelling errors. Communications of the ACM. 1964;7(3):171-6.
84. Meng X-L, Rubin DB. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika. 1993;80(2):267-78.
85. Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, et al. A guide to evaluating linkage quality for the analysis of linked data. International journal of epidemiology. 2017;46(5):1699-710.
86. International Agency for Reseach on Cancer. CI5 – Cancer Incidence in Five Continents Disponible en: http://ci5.iarc.fr/.
87. Pisani P, Bray F, Parkin DM. Estimates of the world-wide prevalence of cancer for 25 sites in the adult population. International journal of cancer. 2002;97(1):72-81.
88. Fiteni F, Westeel V, Pivot X, Borg C, Vernerey D, Bonnetain F. Endpoints in cancer clinical trials. Journal of visceral surgery. 2014;151(1):17-22.
89. Pardo C, Cendales R. Incidencia, mortalidad y prevalencia de cáncer en Colombia, 2012-2016. . Bogotá, Colombia: Instituto Nacional de Cancerología; 2022.
90. Poveda Amaya MC. Prioritizing patients for stomach cancer screening programs: a machine learning approach. Disponible en: https://repository.javeriana.edu.co/handle/10554/62019
91. Leone P, Vacca A, Dammacco F, Racanelli V. Common variable immunodeficiency and gastric malignancies. International journal of molecular sciences. 2018;19(2):451.
92. Charlson M, Szatrowski TP, Peterson J, Gold J. Validation of a combined comorbidity index. Journal of clinical epidemiology. 1994;47(11):1245-51.
93. Charan J, Biswas T. How to calculate sample size for different study designs in medical research? Indian journal of psychological medicine. 2013;35(2):121-6.
94. Pintilie M. Análisis de riesgos competitivos. Revista Española de Cardiología. 2011;64(7):599-605.
95. Lazcano-Ponce E, Martínez ME. Latin America: ripe for cutting-edge research proposals for prevention and control of Helicobacter pylori. Cancer Causes & Control. 2013;24:207-8.
96. Stephens M, Lewis W, White S, Blackshaw G, Edwards P, Barry J, et al. Prognostic significance of alarm symptoms in patients with gastric cancer. Journal of British Surgery. 2005;92(7):840-6.
97. Lee T-Y, Chiang E-P, Shih Y-T, Lane H-Y, Lin J-T, Wu C-Y. Lower serum folate is associated with development and invasiveness of gastric cancer. World Journal of Gastroenterology: WJG. 2014;20(32):11313.

98. Alkan A, Mızrak D, Utkan G. Lower folate levels in gastric cancer: Is it a cause or a result? World journal of gastroenterology. 2015;21(13):4101-2.
99. Zohora F, Bidad K, Pourpak Z, Moin M. Biological and immunological aspects of iron deficiency anemia in cancer development: a narrative review. Nutrition and cancer. 2018;70(4):546-56.
100. Cook MB, Kamangar F, Weinstein SJ, Albanes D, Virtamo J, Taylor PR, et al. Iron in relation to gastric cancer in the Alpha-tocopherol, Beta-carotene Cancer Prevention Study. Cancer epidemiology, biomarkers & prevention. 2012;21(11):2033-42.
101. Rodriguez-Castro KI, Franceschi M, Noto A, Miraglia C, Nouvenne A, Leandro G, et al. Clinical manifestations of chronic atrophic gastritis. Acta Biomed. 2018;89(8-S):88-92.
102. Tang GH, Hart R, Sholzberg M, Brezden-Masley C. Iron deficiency anemia in gastric cancer: a Canadian retrospective review. European journal of gastroenterology & hepatology. 2018;30(12):1497-501.
103. Ford AC, Marwaha A, Sood R, Moayyedi P. Global prevalence of, and risk factors for, uninvestigated dyspepsia: a meta-analysis. Gut. 2015;64(7):1049-57.
104. Carmona-Sánchez R, Gómez-Escudero O, Zavala-Solares M, Bielsa-Fernández M, Coss-Adame E, Hernández-Guerrero A, et al. Consenso mexicano sobre la dispepsia. Revista de Gastroenterología de México. 2017;82(4):309-27.
105. Mahadeva S, Goh K-L. Epidemiology of functional dyspepsia: a global perspective. World journal of gastroenterology: WJG. 2006;12(17):2661.
106. Mahmoodi SA, Mirzaie K, Mahmoodi MS, Mahmoudi SM. A medical decision support system to assess risk factors for gastric cancer based on fuzzy cognitive map. Computational and Mathematical Methods in Medicine. 2020;2020.
107. So JBY, Kapoor R, Zhu F, Koh C, Zhou L, Zou R, et al. Development and validation of a serum microRNA biomarker panel for detecting gastric cancer in a high-risk population. Gut. 2021;70(5):829-37.
108. Rockhill B. The privatization of risk. American journal of public health. 2001;91(3):365.
109. Xie SH, Lagergren J. A model for predicting individuals' absolute risk of esophageal adenocarcinoma: moving toward tailored screening and prevention. International Journal of Cancer. 2016;138(12):2813-9.
110. Lin Z, Wang R, Zhou Y, Wang Q, Yang C-Y, Hao B-C, et al. Prediction of distant metastasis and survival prediction of gastric cancer patients with metastasis to the liver, lung, bone, and brain: Research based on the SEER database. Annals of Translational Medicine. 2022;10(1).
111. van den Boorn HG, Abu-Hanna A, Ter Veer E, van Kleef JJ, Lordick F, Stahl M, et al. SOURCE: a registry-based prediction model for overall survival in patients with metastatic oesophageal or gastric cancer. Cancers. 2019;11(2):187.
112. Pietrantonio F, Miceli R, Raimondi A, Kim YW, Kang WK, Langley RE, et al. Individual patient data meta-analysis of the value of microsatellite instability as a biomarker in gastric cancer. Journal of Clinical Oncology. 2019;37(35):3392-400.
113. Zhu M, Ma Z, Zhang X, Hang D, Yin R, Feng J, et al. C-reactive protein and cancer risk: a pan-cancer study of prospective cohort and Mendelian randomization analysis. BMC medicine. 2022;20(1):1-13.
114. In H, Langdon-Embry M, Gordon L, Schechter CB, Wylie-Rosett J, Castle PE, et al. Can a gastric cancer risk survey identify high-risk patients for endoscopic screening? A pilot study. Journal of Surgical Research. 2018;227:246-56.
115. Stapley S, Peters TJ, Neal RD, Rose P, Walter F, Hamilton W. The risk of oesophago-gastric cancer in symptomatic patients in primary care: a large case–control study using electronic records. British journal of cancer. 2013;108(1):25-31.

116. Pourshams A, Khademi H, Malekshah AF, Islami F, Nouraei M, Sadjadi AR, et al. Cohort profile: the Golestan Cohort Study—a prospective study of oesophageal cancer in northern Iran. International journal of epidemiology. 2010;39(1):52-9.
117. Kim TJ, Pyo JH, Lee H, Baek S-Y, Ahn SH, Min YW, et al. Lack of association between helicobacter pylori infection and various markers of systemic inflammation in asymptomatic adults. The Korean Journal of Gastroenterology. 2018;72(1):21-7.
118. Youssefi M, Tafaghodi M, Farsiani H, Ghazvini K, Keikha M. Helicobacter pylori infection and autoimmune diseases; Is there an association with systemic lupus erythematosus, rheumatoid arthritis, autoimmune atrophy gastritis and autoimmune pancreatitis? A systematic review and meta-analysis study. Journal of Microbiology, Immunology and Infection. 2021;54(3):359-69.
119. Tong L, Wang B-B, Li F-H, Lv S-P, Pan F-F, Dong X-J. An updated meta-analysis of the relationship between Helicobacter pylori infection and the risk of coronary heart disease. Frontiers in Cardiovascular Medicine. 2022;9:794445.
120. Yue L, Zhang R, Chen S, Duan G. Relationship between Helicobacter pylori and incident hypertension as well as blood pressure: a systematic review and meta-analysis. Digestive Diseases. 2023;41(1):124-37.
121. Ezzati Amini E, Moradi Y. Association between helicobacter pylori infection and primary open-angle glaucoma: a systematic review and meta-analysis. BMC ophthalmology. 2023;23(1):374.
122. Haber R, El Gemayel M. Comorbidities in rosacea: a systematic review and update. Journal of the American Academy of Dermatology. 2018;78(4):786-92. e8.
123. Stasi R, Sarpatwari A, Segal JB, Osborn J, Evangelista ML, Cooper N, et al. Effects of eradication of Helicobacter pylori infection in patients with immune thrombocytopenic purpura: a systematic review. Blood, The Journal of the American Society of Hematology. 2009;113(6):1231-40.
124. Wang X, Jiang D, Li T, Zhang X, Wang R, Gao S, et al. Association between microbiological risk factors and neurodegenerative disorders: An umbrella review of systematic reviews and meta-analyses. Frontiers in psychiatry. 2022;13:991085.
125. Shindler-Itskovitch T, Ravona-Springer R, Leibovitz A, Muhsen K. A systematic review and meta-analysis of the association between Helicobacter pylori infection and dementia. Journal of Alzheimer's Disease. 2016;52(4):1431-42.
126. Dabo B, Pelucchi C, Rota M, Jain H, Bertuccio P, Bonzi R, et al. The association between diabetes and gastric cancer: results from the Stomach Cancer Pooling Project Consortium. European Journal of Cancer Prevention. 2022;31(3):260.
127. Tseng C-H. The relationship between diabetes mellitus and gastric cancer and the potential benefits of metformin: An extensive review of the literature. Biomolecules. 2021;11(7):1022.
128. Swanton C, Bernard E, Abbosh C, Andre F, Auwerx J, Balmain A, et al. Embracing cancer complexity: Hallmarks of systemic disease. Cell. 2024;187(7):1589-616.
129. Hasan S. An Overview of Promising Biomarkers in Cancer Screening and Detection. Curr Cancer Drug Targets. 2020;20(11):831-52.
130. Weigt J, Malfertheiner P. Metastatic disease in the stomach. Gastrointestinal tumors. 2015;2(2):61-4.
131. Campoli PMdO, Ejima FH, Cardoso DMM, Silva OQd, Santana Filho JB, Queiroz Barreto PAd, et al. Metastatic cancer to the stomach. Gastric Cancer. 2006;9:19-25.
132. Namikawa T, Marui A, Yokota K, Kawanishi Y, Munekage M, Uemura S, et al. Frequency and therapeutic strategy for patients with ovarian metastasis from gastric cancer. Langenbeck's Archives of Surgery. 2022;407(6):2301-8.
133. Pardo Ramos C, De Vries E, Buitrago L, Gamboa O. Atlas de mortalidad por cáncer en colombia. Cuarta edición ed: Instituto Nacional de Cancerología; 2017.

134. Smith RA. The value of modern mammography screening in the control of breast cancer: understanding the underpinnings of the current debates. Cancer epidemiology, biomarkers & prevention. 2014;23(7):1139-46.
135. Mizota Y, Yamamoto S. How long should we continue gastric cancer screening? From an epidemiological point of view. Gastric Cancer. 2019;22:456-62.
136. Yang Y, Long Z, Zhong Z, Liu Q, Yang X. Construction and Evaluation of Gastric Cancer Risk Prediction Model. Indian Journal of Pharmaceutical Sciences. 2021;83:112-8.
137. Zhu X, Lv J, Zhu M, Yan C, Deng B, Yu C, et al. Development, validation, and evaluation of a risk assessment tool for personalized screening of gastric cancer in Chinese populations. BMC medicine. 2023;21(1):159.
138. Cao M, Li H, Sun D, Lei L, Ren J, Shi J, et al. Classifying risk level of gastric cancer: Evaluation of questionnaire-based prediction model. Chinese Journal of Cancer Research. 2020;32(5):605.
139. Watanabe J, Kotani K. The effect of Helicobacter pylori eradication on C-reactive protein: results from a meta-analysis. Archives of Medical Science: AMS. 2022;18(4):958.
140. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence. 2016;5(4):221-32.
141. Steinberg DM, Fine J, Chappell R. Sample size for positive and negative predictive value in diagnostic research using case-control designs. Biostatistics. 2009;10(1):94-105.