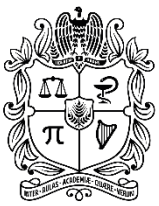


UNIVERSIDAD
NACIONAL
DE COLOMBIA

Modelo farmacogenético y clínico para la predicción de desenlaces en pacientes con artritis reumatoide tratados con metotrexato y adalimumab

Fabián Alberto Hernández Tarapués

Universidad Nacional de Colombia

Facultad de ingeniería, Departamento de Sistemas e Industrial

Bogotá, Colombia

2020

Modelo farmacogenético y clínico para la predicción de desenlaces en pacientes con artritis reumatoide tratados con metotrexato y adalimumab

Fabián Alberto Hernández Tarapués

Tesis presentada como requisito parcial para optar al título de:

Magister en Bioinformática

Director:

Ph.D., Luis Fernando Niño Vásquez

Codirector:

Ph.D., Fabio Ancízar Aristizábal Gutiérrez

Línea de Investigación:

Farmacogenética

Grupo de Investigación:

Laboratorio de Investigación en Sistemas Inteligentes – LISI

Universidad Nacional de Colombia

Facultad de ingeniería, Departamento de Sistemas e Industrial

Bogotá, Colombia

2020

“The one prudence in life is concentration; the one evil is dissipation; and it makes no difference whether our dissipations are coarse or fine; property and its cares, friends and a social habit, or politics, or music, or feasting. Everything is good which takes away one plaything and delusion more and drives us home to add one stroke of faithful work.”

Ralph Waldo Emerson

Agradecimientos

Me gustaría agradecer a Patricia Zuluaga, estudiante de Maestría en Farmacología por el acompañamiento y camaradería en la recolección de información del Hospital Militar, a los doctores John Londoño y Jesús Ballesteros por su ayuda en los trámites administrativos para someter el proyecto en el Hospital Militar Central y sus aportes desde la experiencia clínica que permitieron guiar la concepción y realización del proyecto.

Adicionalmente, agradezco especialmente a Viviana Ariza por su amistad y apoyo moral, cuando parecía que este trabajo nunca iba a terminar y a los integrantes del grupo LISI por sus acertados aportes durante todo el desarrollo de este trabajo.

Resumen

OBJETIVO: Desarrollar un modelo farmacogenético y clínico para la predicción de desenlaces de efectividad en una cohorte de pacientes diagnosticados con artritis reumatoide (AR) tratados con metotrexato o adalimumab en el Hospital Militar Central.

MÉTODOS: Se probaron cinco métodos de aprendizaje automático en el conjunto de datos con previo preprocesamiento para limpieza y selección de variables: Regresión logística, árboles de decisión, bosques aleatorios, máquinas de soporte vectorial (SVM) y redes neuronales artificiales (ANN) en una cohorte de 155 pacientes tratados con MTX que fue derivada en una cohorte de entrenamiento (124 pacientes) y una de prueba (31 pacientes).Se incluyeron tanto variables clínicas como variaciones genéticas El desenlace escogido fue la respuesta a la terapia medida como un puntaje DAS 28 < 3,2. El criterio de evaluación de desempeño fue el área bajo la curva (AUC) de las características operativas del receptor (ROC).

RESULTADOS: Los algoritmos con mayor poder predictivo fueron las SVM y las ANN. Las principales variables seleccionadas para la cohorte de MTX fueron la edad, tiempo con AR, clasificación funcional y genotipos de las variantes rs9344, rs4148396, rs4673993, rs1801133 y rs7279445. Dado el tamaño de la cohorte de pacientes tratados con ADA (12 pacientes), no se pudo ajustar de forma exitosa ningún modelo de aprendizaje automático.

CONCLUSIONES: Se desarrolló un modelo pronóstico con un poder predictivo alto en la cohorte de pacientes tratados con MTX que identifica pacientes propensos a no responder al tratamiento.

Palabras clave: Artritis Reumatoide, Farmacogenética, Modelo Predictivo, Aprendizaje Automático.

Abstract

GOAL: To develop a pharmacogenetic and clinical model to predict effectiveness outcomes in a cohort of patients diagnosed with rheumatoid arthritis (RA) treated with methotrexate or adalimumab at the Central Military Hospital in Bogota, Colombia.

METHODS: Five statistical learning methods were tested on the data set with previous pre-processing for variable cleaning and selection: Logistic regression, decision trees, random forests, Support Vector Machines (SVM) and Artificial Neural Networks (ANN). The models were applied in a cohort of 155 patients treated with MTX which was derived in a training (124 patients) and a test cohort (31 patients). Both clinical variables and genetic variations were included. The chosen outcome was the therapy response measured as a DAS 28 score <3.2. The performance evaluation criterion was the area (AUC) under the receiver operating characteristics (ROC) curve.

RESULTS: The algorithms with the highest predictive power were SVM and ANN. For the MTX cohort, the main selected variables were age, time with RA, functional classification, and genotypes of the rs9344, rs4148396, rs4673993, rs1801133 and rs7279445 variants. Given the size of the cohort of ADA-treated patients (12 patients), no machine learning model could be successfully adjusted.

CONCLUSIONS: A prognostic model with high predictive power was developed in the cohort of patients treated with MTX, which is able to identify patients prone to not responding well to treatment.

Keywords: Rheumatoid Arthritis, Pharmacogenetics, Predictive Model, Machine Learning

Contenido

Pág.

Resumen ..... **IX**

Lista de figuras ..... **XII**

Lista de tablas ..... **XIV**

Lista de abreviaturas ..... **XVI**

Introducción **1**

1. Artritis Reumatoide **3**

2. Modelos de aprendizaje automático en farmacogenómica **13**

3. Métodos de selección de variables **31**

4. Revisión de métodos para su aplicación en conjuntos de datos biomédicos **38**

5. Metodología **51**

6. Resultados **69**

7. Discusión **109**

8. Conclusiones y trabajo futuro **113**

9. Anexos **115**

10. Referencias **160**

Lista de figuras

Pág.

Figura 1. Factores que influyen en la eficacia y seguridad de un fármaco. Tomada de (7) 11
Figura 2. Representación del hiperplano en el espacio de p dimensiones como una línea en un espacio de 2 dimensiones. La margen del hiperplano se muestra como una línea punteada..... 22
Figura 3. Representación de un conjunto de puntos que no pueden separarse perfectamente por un hiperplano. Izquierda: Distribución de los puntos. Derecha: Clasificador de soporte vectorial que permite la localización errónea de algunos puntos con respecto al hiperplano. De esta manera, se observan puntos que están en el lado correcto de la margen y del hiperplano, (a) en la margen del hiperplano, (b) en el lado correcto del hiperplano, pero en el lado incorrecto de la margen y (c) en el lado incorrecto del hiperplano. 24
Figura 4. Representación de un problema no separable linealmente. Se observa la generación de una dimensión adicional para permitir la separación lineal de un problema que inicialmente no podría ser separable de forma lineal..... 26
Figura 5. Esquema de funcionamiento de una neurona dentro de una red neuronal artificial 28
Figura 6. Análisis de correspondencias múltiples del conjunto de datos de la Tabla 3 ... 35
Figura 7. Diagrama PRISMA con los resultados de la revisión..... 43
Figura 8. Esquema de genotipificación con MassARRAY®, cada corrida se va almacenando para dar la gráfica de la derecha. Los puntos rojos esquematizan aquellos que no pudieron ser agrupados en el análisis. Tomado de (176). 59
Figura 9. Significado de las nubes de puntos en el llamado de bases. Los tres grupos de genotipos representan las nubes generadas a partir de gráficos de intensidad. El grupo AA consta de todas las llamadas homocigóticas menores, las llamadas heterocigotas corresponden al grupo AB y las llamadas homocigóticas mayores corresponden al grupo BB. Tomado de (177)..... 59
Figura 10. Posiciones relativas y resultados del análisis de desequilibrio de ligamiento para los SNPs considerados para MTX..... 72
Figura 11. Posiciones relativas y resultados del análisis de desequilibrio de ligamiento para los SNPs considerados para ADA..... 73
Figura 12. Porcentaje de varianza explicado por cada dimensión en el ACM para las variables clínicas..... 78

Figura 13.Contribución relativa de las variables en la construcción de las dos primeras dimensiones para las variables clínicas 79
Figura 14. Primeras dos dimensiones del análisis y posición de las variables clínicas en cada una 80
Figura 15. Posición de los individuos en las dos primeras dimensiones para el ACM de las variables clínicas 81
Figura 16. Distribución de las etiquetas "respondedor" y "no respondedor" de cada individuo en las dos dimensiones generadas con las variables clínicas 82
Figura 17. Porcentaje de varianza explicado por cada dimensión en el ACM para las variables genéticas..... 83
Figura 18. Contribución relativa de las variables en la construcción de las dos primeras dimensiones para las variables genéticas 83
Figura 19. Izquierda: Primeras dos dimensiones del análisis y posición de las variables genéticas en cada una. Derecha: Posición de los individuos en las dos primeras dimensiones para el ACM de las variables genéticas..... 84
Figura 20. Distribución de los genotipos de los SNPs rs4673993 y rs2372536 de cada individuo en las dos dimensiones generadas con las variables genéticas 84
Figura 21. Importancia ordenada de las variables clínicas para el conjunto de MTX 87
Figura 22. Importancia ordenada de las variables genéticas para el conjunto de MTX.. 88
Figura 23. Resultados del proceso de eliminación recursiva de características utilizando los bosques aleatorios como algoritmo base. Izquierda: Variables clínicas. Derecha: Variables genéticas 89
Figura 24. Resultados del proceso de eliminación recursiva de características utilizando *Naive Bayes* como algoritmo base. Izquierda: Variables clínicas. Derecha: Variables genéticas..... 89
Figura 25. Resultados del proceso de eliminación recursiva de características utilizando árboles de decisión con *bagging* como algoritmo base. Izquierda: Variables clínicas. Derecha: Variables genéticas..... 90
Figura 26. Importancia ordenada de las variables genéticas para el conjunto de ADA .. 94
Figura 27. Curvas ROC con validación cruzada para la regresión logística..... 97
Figura 28. Curvas ROC con validación cruzada para los árboles de decisión 100
Figura 29. Tasa de error (*misclassification)* como función de las iteraciones de *AdaBoost* 101
Figura 30. Curvas ROC con validación cruzada para los árboles de decisión con *AdaBoost*.....102
Figura 31. Errores de clasificación en el bosque como función del número de árboles que lo componen.....103
Figura 32. Curvas ROC con validación cruzada para los bosques aleatorios 104
Figura 33. Curvas ROC para las SVM 106
Figura 34. Curvas ROC para las ANN 107

Lista de tablas

Pág.

Tabla 1. Criterios de clasificación ACR/EULAR para AR. Tomado de (40). 4
Tabla 2. Cálculo y puntos de corte para la puntuación DAS28. Tomado de (54). 7
Tabla 3. Conjunto de datos de ejemplo con variables categóricas relacionadas con características físicas de un grupo de individuos 34
Tabla 4. Resultados de la evaluación de pertinencia de las publicaciones con respecto a las preguntas. Resaltadas se encuentran las publicaciones descartadas de la síntesis.. 43
Tabla 5. Resultados de la prueba chi-cuadrado de bondad de ajuste para el equilibrio de Hardy-Weinberg. "-"Implica que no había representación de algún genotipo en el conjunto de datos, por lo tanto las frecuencias no podían evaluarse. 70
Tabla 6. Resultados del análisis asociativo univariado con los diferentes modelos genéticos para los SNPs considerados para MTX. "-"Implica que no había representación de algún genotipo en el conjunto de datos, por lo tanto las frecuencias no podían evaluarse 74
Tabla 7. Resultados del análisis asociativo univariado con los diferentes modelos genéticos para los SNPs considerados para ADA. "-"Implica que no había representación de algún genotipo en el conjunto de datos, por lo tanto las frecuencias no podían evaluarse 75
Tabla 8. Resultado de la prueba de chi-cuadrado para el conjunto de datos de MTX..... 85
Tabla 9. Top 10 de variables clínicas seleccionadas por mRMR 85
Tabla 10. Top 10 de variables genéticas seleccionadas por mRMR 86
Tabla 11. Variables seleccionadas con cada algoritmo..... 90
Tabla 12. Coeficientes significativos para la regresión logística de variables clínicas y genéticas 91
Tabla 13. Coeficientes obtenidos en la regresión logística final 95
Tabla 14. Matriz de confusión para la regresión logística 96
Tabla 15. Resultados de la validación cruzada para la regresión logística..... 96
Tabla 16. Matriz de confusión para el árbol de decisión 99
Tabla 17. Resultados de la validación cruzada para el árbol de decisión..... 99
Tabla 18. Matriz de confusión para el árbol de decisión con *boosting* 100
Tabla 19. Resultados de la validación cruzada de los árboles de decisión con *AdaBoost* 101
Tabla 20. Matriz de confusión para los bosques aleatorios..... 103
Tabla 21. Resultados de la validación cruzada para los bosques aleatorios..... 104

Tabla 22. Matriz de confusión para las SVM 105
Tabla 23. Resultados de la validación cruzada para las SVM 105
Tabla 24. Resultados de la validación cruzada para las ANN 107

Lista de abreviaturas

Abreviatura	Término
ABC	ATP-Binding Cassette
ACM	Análisis de Correspondencias Múltiples
ADA	Adalimumab
AIC	Criterio de Información de Akaike
ANN	Redes Neuronales Artificiales – *Artificial Neural Networks*
AR	Artritis Reumatoide
AUC	Área Bajo la Curva – *Area Under the Curve*
CART	Árboles de Regresión y Clasificación
DAS28	Disease Activity Score de 28 articulaciones
DE	Desviación Estándar
DNA	Ácido Desoxirribonucleico – *Deoxyribonucleic Acid*
EHW	Equilibrio de Hardy-Weinberg
EULAR	Liga Europea Contra el Reumatismo
FARME	Fármacos Antirreumáticos Modificadores de la Enfermedad
GPC	Guía de Práctica Clínica
IL	Interleucina
MMC	Clasificador de Margen Máxima
mRMR	Máxima Relevancia y Mínima Redundancia
MTX	Metotrexato
OOB	Fuera de la bolsa – *Out of Bag*

Abreviatura	Término
OR	Razón de momios – *Odds Ratio*
PCR	Proteína C Reactiva
RFE	Eliminación Recursiva de Características – *Recursive Feature Elimination*
RNA	Ácido Ribonucleico – *Ribonucleic Acid*
ROC	Características Operativas del Receptor – *Receiver Operating Characteristic*
RSS	Residuo de la Suma de Cuadrados – *Residual Sum of Squares*
SNP	Polimorfismo de Único Nucleótido – *Single Nucleotide Polymorphism*
SVM	Máquinas de Soporte Vectorial – *Support Vector Machines*
TNF	Factor de Necrosis Tumoral – *Tumor Necrosis Factor*

Introducción

La artritis reumatoide (AR) es una enfermedad crónica, inflamatoria que afecta principalmente las articulaciones y reduce la movilidad del individuo (1). El tratamiento de esta enfermedad consiste usualmente en inmunosupresores, analgésicos y fisioterapia, con el fin de conservar la funcionalidad y disminuir la probabilidad de aparición de comorbilidades sistémicas (2). Dentro de los tratamientos más utilizados se encuentran el metotrexato y los inhibidores del factor de necrosis tumoral alfa, cuyo mecanismo de acción principal consiste en retrasar el proceso inflamatorio y de autoinmunidad (3). Sin embargo, un porcentaje importante de los pacientes con RA no responden adecuadamente al tratamiento farmacológico (4), lo que tiene un impacto directo en el logro de las metas terapéuticas y, en últimas, en su calidad de vida (5).

Una de las hipótesis más aceptadas en cuanto a la variabilidad en la respuesta a tratamientos farmacológicos para la AR es la de la heterogeneidad implícita de la enfermedad, debido a la compleja interacción de factores bioquímicos, ambientales y fisiológicos que desembocan en diferentes niveles de erosión de cartílago y hueso, aparición de complicaciones multisistémicas y sensibilidad a inmunomoduladores (6–9).

En respuesta a esta hipótesis, muchas aproximaciones de medicina personalizada en AR se han desarrollado en los últimos años (9,10), con el fin de establecer características particulares del individuo que puedan ser las fuentes de la heterogeneidad tanto en la presentación de la enfermedad como en la respuesta a tratamientos farmacológicos.

El paradigma de la medicina personalizada ha reemplazado en algunos escenarios al enfoque tradicional de la medicina, donde los pacientes son estratificados en grupos de severidad o presentación de acuerdo con recomendaciones de guías de práctica clínica (GPC) para luego escoger el tratamiento que mejor se adecue a ese grupo particular de pacientes (3,4).

Dentro de las aproximaciones relacionadas con la medicina personalizada, la farmacogenómica ha sido una de las más utilizadas, debido a la reciente explosión de datos ómicos (11). La farmacogenómica está definida como el estudio de las variaciones genéticas y su relación con la variabilidad de la respuesta a tratamientos farmacológicos (7). En este aspecto, se han realizado diversos estudios de asociación de genoma completo, con el fin de detectar las variaciones genéticas que estarían relacionadas con la variabilidad en la respuesta a tratamientos para la AR (12,13), así como otros estudios de asociación genética para establecer la correlación entre polimorfismos en determinados genes y la respuesta a tratamientos como el metotrexato (14–16) y los inhibidores del factor de necrosis tumoral alfa (17–20).

Con el fin de considerar la contribución tanto de las variables genéticas como de los parámetros clínicos del paciente, se han desarrollado distintas aproximaciones de integración de datos con fines predictivos (21–25) utilizando diversos algoritmos de aprendizaje automático, que pueden considerar tanto los efectos compensadores y redundancia de las diferentes variables, dado que diferentes estudios de asociación univariada han arrojado resultados contradictorios para las mismas variables genéticas en pacientes diagnosticados con AR (26–31).

1. Artritis Reumatoide

La artritis reumatoide (AR) es una enfermedad crónica, inflamatoria y sistémica, caracterizada por una afectación persistente y crónica de las articulaciones, que puede llevar al daño del cartílago y hueso, así como a la discapacidad progresiva y complicaciones de carácter sistémico (32). La AR es una enfermedad heterogénea con una variedad de presentaciones clínicas y mecanismos de patogenicidad involucrados, incluso en individuos con el mismo tipo de diagnóstico o en diferentes estadios de la enfermedad (33).

1.1 Fisiopatología y clasificación

Aunque el mecanismo de aparición de la AR aún no ha sido completamente entendido, se ha observado que existe una compleja interacción de factores genéticos y ambientales que podría llevar al desarrollo de la enfermedad (34). Así, en los individuos genéticamente predispuestos, los factores ambientales como la obesidad, dieta, fumar, microbiota gastrointestinal e infecciones podrían desencadenar la activación aberrante de la respuesta inmune innata y adaptativa, causando la pérdida de tolerancia inmunológica, presentación de autoantígenos con activación de células T y B, y producción de citoquinas inflamatorias (32,35). Esta cascada de eventos es lo que eventualmente lleva a la sinovitis, destrucción de cartílago y hueso y otros síntomas extraarticulares característicos de la enfermedad (32,33).

La patogénesis de la AR comienza en la sinovia, que es una estructura entre las dos superficies cartilaginosas de la articulación. Esta estructura está encargada de proveer nutrientes al cartílago (que no tiene suministro propio de sangre) y producir lubricantes para disminuir la fricción entre estas dos superficies (1). La activación aberrante de la respuesta inmune resulta en la activación de los sinoviocitos, que hacen que esta estructura se expanda y comience a producir citoquinas proinflamatorias como la

interleucina (IL)-1, IL6, factor de necrosis tumoral (TNF), entre otros (36,37). Posteriormente, se da la infiltración de otro tipo de células inmunes a esta estructura, como las células T y B, que junto con macrófagos, mastocitos y neutrófilos contribuyen a la degeneración del cartílago y hueso (1,38).

Teniendo en cuenta la heterogeneidad de la enfermedad, el proceso de diagnóstico de la AR es un proceso altamente individualizado, ya que no existen criterios diagnósticos, sino criterios de clasificación de la enfermedad, como ocurre en muchas otras patologías reumáticas (34). Los criterios de clasificación tienen como objetivo estratificar pacientes con características similares para fines de investigación clínica, pero no tienen la intención de capturar a todos los pacientes con una alta especificidad; sin embargo, debido a la falta de criterios diagnósticos, estos son frecuentemente utilizados en la práctica clínica y en la toma de decisiones en salud (39). De esta manera, con base en la fisiopatología de la enfermedad, los síntomas más evidentes y utilizados en el diagnóstico son la inflamación y sensibilidad de las articulaciones, que pueden ser complementados con exámenes serológicos (como el factor reumatoideo – FR o los anticuerpos anti-péptidos citrulinados cíclicos – ACPA) y reactantes de fase aguda (proteína C reactiva – PCR o velocidad de sedimentación globular – VSG) (34). Así, se han creado criterios de clasificación que tienen en cuenta las variables anteriormente mencionadas y han sido ampliamente utilizados tanto en la investigación como en la práctica clínica, como lo son los criterios del Colegio Americano de Reumatología (ACR por sus siglas en inglés) y de la Liga Europea contra el Reumatismo (EULAR por sus siglas en inglés) (40), que se resumen en la ***Tabla 1***.

Tabla 1. Criterios de clasificación ACR/EULAR para AR. Tomado de (40).

Descripción	**Puntuación**
Población objetivo: 1). Pacientes que tengan al menos una articulación con sinovitis clínicamente definida. 2). Pacientes con sinovitis que no se pueda explicar por alguna otra patología.	
A). Articulaciones involucradas 1 articulación grande (hombros, codos, caderas, rodillas, caderas y tobillos)	 0 1

2-10 articulaciones grandes	2
1-3 articulaciones pequeñas (muñecas y falanges)	3
4-10 articulaciones pequeñas	5
>10 articulaciones (con al menos 1 articulación pequeña)	
B). Serología (al menos un examen es necesario)	
FR negativo y ACPA negativo	0
FR positivo-bajo o ACPA positivo-bajo	2
FR positivo-alto o ACPA positivo-alto	3
C). Reactantes de fase aguda	
PCR normal y VSG normal	0
PCR anormal o VSG anormal	1
D). Duración de los síntomas	
< 6 semanas	0
≥ 6 semanas	1
Un resultado mayor o igual a 6/10 indica un diagnóstico de Artritis Reumatoide	

.

1.2 Epidemiología y carga de la enfermedad

De acuerdo con diversas estimaciones, la prevalencia de la AR en Colombia varía de 0,2% a 0,9% (41–43) y la incidencia es de 10,4 por cada 100.000 habitantes (43). Consistentemente con reportes de otras latitudes, esta enfermedad se presenta mayoritariamente en mujeres, con una relación mujer hombre de 4,2:1 para Colombia (42).

Al ser una enfermedad musculoesquelética degenerativa, la artritis reumatoide tiene un alto impacto socioeconómico y en la calidad de vida del individuo (5). De acuerdo con un estudio realizado en Suecia, alrededor de tres cuartos del costo total que representa la AR está relacionado con la pérdida de productividad laboral (44). Adicionalmente, de acuerdo con los resultados del estudio QUEST-RA (*Quantitative Standard Monitoring of Patients with Reumathoid Arthritis*) que analizó la productividad laboral y la discapacidad relacionada con la AR en más de 8000 pacientes de 16 países con producto interno bruto

(PIB) alto (mayor a 24.000 USD per cápita) y 16 países con PIB bajo (menor a 11.000 USD per cápita), entre los que se encontraban dos países de América Latina, el 37% de los pacientes reportó incapacidad laboral debido a la AR. Más aún, en los países con bajo PIB la severidad de la enfermedad era mayor en los individuos que continuaban trabajando, comparado con los individuos de países con PIB alto que dejaban de trabajar (45). Esta información indica que la pérdida de la productividad laboral junto con la discapacidad progresiva que genera la enfermedad representan una carga importante desde la perspectiva de la sociedad, que podría disminuirse con intervenciones que mejoren la función motora y retrasen la progresión de la enfermedad en los individuos.

Finalmente, la AR está clasificada en Colombia como una enfermedad de alto costo, donde la principal fuente de elevación de costos reside en los medicamentos biológicos indicados para algunos pacientes (43). En un estudio farmacoepidemiológico realizado en una cohorte de más de 1’500.000 pacientes afiliados al sistema de salud colombiano, se encontró que el 4,4% de los afiliados estaba consumiendo algún medicamento antirreumático y de estos el 4,6% de los pacientes estaba en tratamiento con algún inhibidor del TNF-α; sin embargo, el costo de tratamiento mensual de estos pacientes ascendía a más de 30 millones de pesos colombianos (COP) del 2009 (46). Por otra parte, en un estudio realizado en pacientes con AR temprana de un hospital de Bogotá, se encontró que un paciente con actividad de la enfermedad leve durante el primer año cuesta COP $3’325.796, con actividad moderada $COP 3’554.193 y con actividad grave COP $46’155.596 del 2008; de estos costos el 86% correspondió al costo de medicamentos, 10% al costo de laboratorios y 4% a la atención médica. Cabe agregar que de los tres grupos de pacientes considerados, sólo el último recibía tratamiento con medicamentos biológicos (47). Estos resultados son consistentes con los obtenidos en un estudio similar realizado en la ciudad de Medellín en 2011, donde el costo de tratamiento de pacientes que recibían terapia biológica era 7,6 veces mayor que el costo de tratamiento de pacientes que no la recibían (48).

Así, la discapacidad progresiva, disminución de la productividad laboral y alto costo que representa para el sistema de salud, hacen de esta enfermedad un problema importante de salud pública, donde el desarrollo de herramientas para un adecuado uso de recursos en salud como los medicamentos se vuelva relevante.

1.3 Tratamiento farmacológico y farmacogenética

Si bien la AR es, hasta el momento, incurable, las aproximaciones terapéuticas modernas han cambiado radicalmente el paradigma de tratamiento de la enfermedad, permitiendo la consecución del control de la enfermedad y un mejoramiento dramático en los desenlaces (33). En general, las estrategias actuales de tratamiento de la AR involucran una aproximación *treat-to-target* basado en la monitorización estrecha y continua de la actividad de la enfermedad, con el fin de basar la toma de decisiones de manejo terapéutico en objetivos que implican la mayor disminución posible de la actividad de la enfermedad (49–51). Esto busca que la consecución de los objetivos se traduzca en una mejora de la función motora, disminuir la ocurrencia de daño óseo y prevención de la destrucción de la articulación (52).

Por otra parte, teniendo en cuenta la heterogeneidad de la enfermedad, no existe una medida única que permita establecer su severidad o la efectividad de los tratamientos; por lo tanto, se han establecido puntuaciones que tienen en cuenta una variedad de biomarcadores, con el fin de tener en cuenta todos los componentes que tienen injerencia en la actividad de la enfermedad (53). Dentro de las puntuaciones más utilizadas en la práctica clínica se encuentra el *Disease Activity Score* (DAS) y su modificación que utiliza el conteo de 28 articulaciones (DAS28) (53,54). Esta puntuación pondera el conteo de articulaciones inflamadas y dolorosas, los reactantes de fase aguda (VSG o PCR) y un examen global general del médico. En general, en esta puntuación se da mayor peso a los reactantes de fase aguda que al conteo de articulaciones, por lo tanto, las terapias que tengan alto impacto en VSG o PCR serán clasificadas como muy efectivas por esta puntuación (55). El cálculo de esta puntuación, junto con los puntos de corte de evaluación se muestran en la **Tabla 2**.

Tabla 2. Cálculo y puntos de corte para la puntuación DAS28. Tomado de (54).

Fórmula $DAS28 = 0,56 \times \sqrt{28TJC} + 0,28 \times \sqrt{28SJC} + 0,70 \times \ln VSG + 0,014 \times GH$	
Puntuación	**Significado**
DAS28 < 2,6	Periodo de remisión de la enfermedad

2,6 ≤ DAS28 ≤ 3,2	Baja actividad de la enfermedad
3,2 < DAS28 ≤ 5,1	Actividad moderada de la enfermedad
DAS28 > 5,1	Alta actividad de la enfermedad
Donde 28TJC es el conteo de articulaciones dolorosas, 28SJC es el conteo de articulaciones inflamadas, ESR es la velocidad de sedimentación globular y GH es una puntuación del estado de salud en general del individuo. Un cambio en el DAS28 ≥ 1,2 es considerado un cambio significativo, debido a que es muy improbable que sea el resultado de un error de medida aleatorio (P ≤ 0,05)(56)	

1.3.1 Tratamiento con FARMEs

Los pacientes con diagnóstico de AR son tratados generalmente con Fármacos Antirreumáticos Modificadores de la Enfermedad (FARMEs), definidos como moléculas que interfieren con los signos y síntomas de la enfermedad, mejoran la movilidad y retrasan la progresión y el daño articular (34). Los FARMEs están clasificados en moléculas de síntesis química y biológicos: los primeros son moléculas pequeñas que generalmente son administradas de forma oral y los segundos corresponden a proteínas terapéuticas administradas de forma parenteral (57).

Entre los FARMEs de síntesis química más utilizados en el tratamiento de la AR se encuentra el metotrexato (MTX), que ha sido usado en AR desde hace más de 50 años (58). Sin embargo, a pesar de su uso extendido en la práctica clínica, el mecanismo por el cual el MTX ejerce su efecto modificador de la enfermedad todavía no se encuentra bien elucidado, lo que contrasta con el detallado entendimiento de su mecanismo de acción como terapia antineoplásica (59). Se sabe que el efecto antirreumático de este fármaco debería tener mecanismos alternativos al que surte cuando ejerce su acción antineoplásica, debido a que la administración de su antídoto en quimioterapia, el folato, no ha demostrado disminuir el beneficio terapéutico del fármaco en AR (60). Se cree que los metabolitos poliglutamados del MTX son los responsables del efecto farmacológico de la molécula en AR, teniendo en cuenta que existe una relación de temporalidad entre el comienzo de la respuesta y la aparición de concentraciones importantes de los metabolitos mencionados (61). En Colombia, de acuerdo con la Guía de Práctica Clínica (GPC) para

la detección temprana, diagnóstico y tratamiento de la artritis reumatoide, se recomienda el abordaje terapéutico *treat-to-target* en pacientes diagnosticados con AR, así como el inicio de la terapia con FARMEs de síntesis química, dentro de los cuales existe evidencia fuerte a favor de iniciar el tratamiento con MTX (62). El MTX puede ser administrado en combinación con otros agentes no modificadores de la enfermedad, como antiinflamatorios y analgésicos, o con glucocorticoides por un periodo limitado de tiempo.

Cuando el tratamiento con MTX (con o sin otros agentes no modificadores de la enfermedad) no ejerce un efecto terapéutico suficiente, los pacientes son generalmente estratificados de acuerdo con sus factores de riesgo para recibir un FARME sintético adicional o un FARME biológico o un FARME sintético dirigido (como tofacitinib o baricitinib) (62,63). Dentro de los FARMEs biológicos hay una gran variedad de mecanismos de acción: la inhibición de CD80 y CD86 (abatacept), la inhibición del receptor IL-6 (tocilizumab y sarilumab), la inhibición de CD20 (rituximab) y la inhibición del TNF-α (adalimumab, certolizumab pegol, etanercept, infliximab, golimumab e infliximab) (33,34). En general, todos los biológicos son más eficaces cuando están combinados con FARMEs de síntesis química que en monoterapia únicamente y por tanto, se habla de adición de biológicos a MTX cuando la respuesta a este último es insuficiente (64–68).

Adalimumab (ADA) es un anticuerpo IgG1 recombinante totalmente humanizado, cuyo mecanismo de acción es la unión específica al TNF-α, neutralizando la actividad de esta citoquina. Es un medicamento que se administra de forma subcutánea (69). En Colombia está recomendado como tratamiento adicional a FARMEs de síntesis cuando el paciente ha fallado a monoterapia o terapia combinada con estos agentes (62).

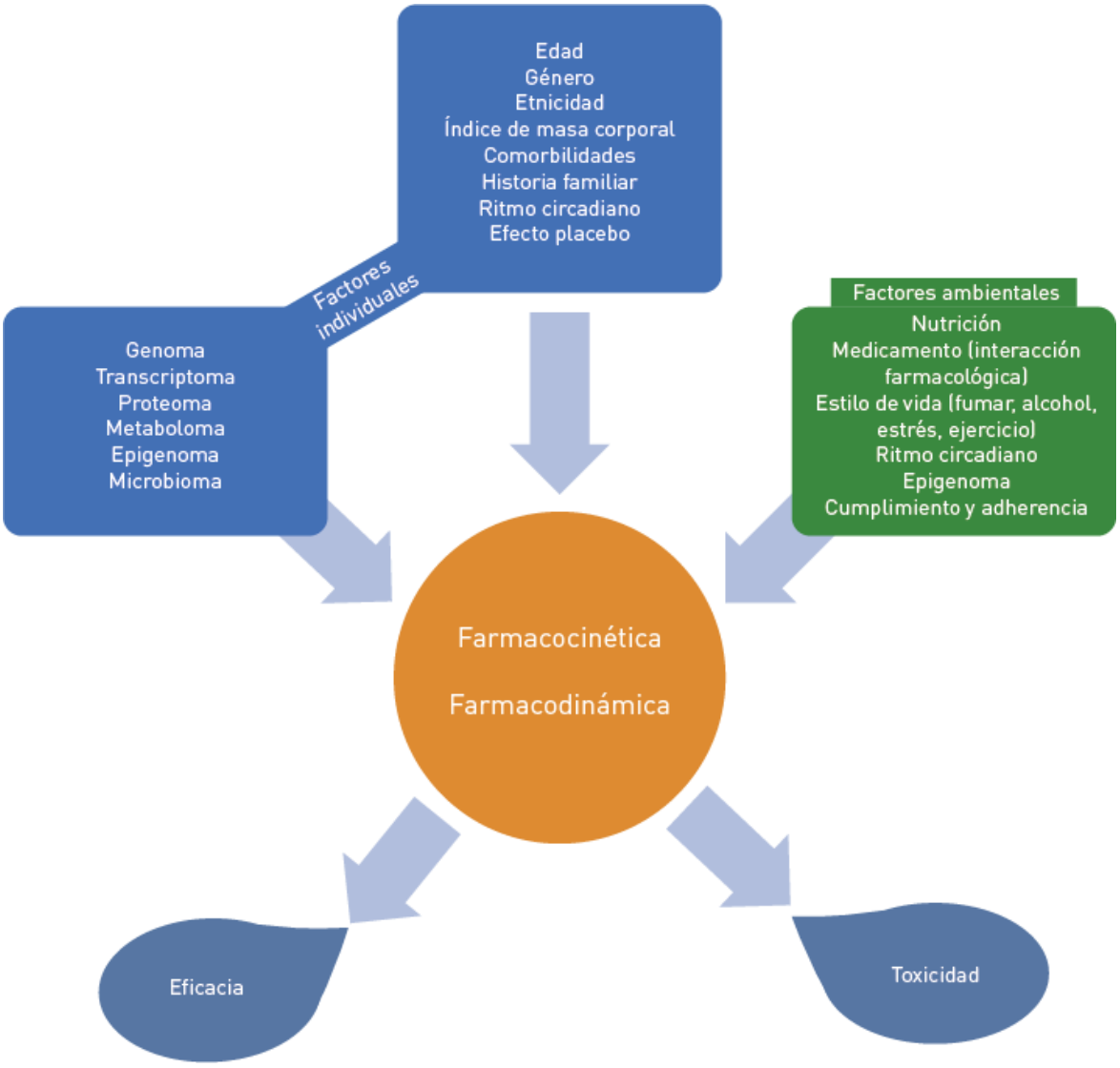
La elección de la terapia antirreumática se realiza teniendo en cuenta la condición del paciente y estadio de la enfermedad, atendiendo también a las recomendaciones de las guías de práctica clínica (3,4). Sin embargo, esta aproximación ha demostrado no ser del todo efectiva, teniendo en cuenta que la remisión es lograda por el 5% al 43% de los pacientes tratados con FARMEs, de acuerdo con los resultados de diferentes estudios (50,70–72).

1.3.2 Farmacogenética

Dado el frecuente problema de la variabilidad interindividual en la respuesta a terapias antirreumáticas, se acude cada vez más frecuentemente al uso de la farmacogenética.

Definida como el estudio de las variantes genéticas y su relación con los desenlaces de las terapias (7), la farmacogenética ha sido utilizada como una aproximación para el entendimiento de los mecanismos subyacentes involucrados en la variabilidad de la respuesta a fármacos y por tanto, como herramienta predictiva para los desenlaces terapéuticos en cuanto a eficacia y seguridad en enfermedades altamente heterogéneas, como la AR (6,73,74), teniendo en cuenta que estos desenlaces son el resultado de una compleja interacción entre factores ambientales, genéticos y epigenéticos que influyen, en última instancia, en la farmacocinética y farmacodinamia de los medicamentos (**Figura 1**) (7). La utilización de información de fuentes ómicas (genómica, transcriptómica, proteómica, metabolómica, entre otras) en este tipo de estudios de asociación se denomina farmacogenómica (7). Numerosos estudios farmacogenéticos han sido realizados con el fin de establecer asociaciones entre polimorfismos genéticos y desenlaces de las terapias con FARMEs en cuanto a efectividad clínica y seguridad (26–31).

Figura 1. Factores que influyen en la eficacia y seguridad de un fármaco. Tomada de (7)



Los estudios de asociación de genoma completo (GWAS por sus siglas en inglés: *Genome-Wide Association Studies*) y un mayor entendimiento de la dinámica intracelular del MTX (12,75) han permitido establecer una lista de posibles variantes genéticas implicadas en la heterogeneidad en la respuesta a MTX (tanto en eficacia como en seguridad) y corresponden, en su mayoría a polimorfismos en genes que codifican proteínas involucradas en la farmacocinética y farmacodinamia de la molécula (6,75). Por otra parte, diferentes estudios han evaluado la relación entre polimorfismos genéticos y la respuesta a tratamientos con inhibidores del TNF-α (76–79). De todos los polimorfismos estudiados, la evidencia ha permitido establecer que polimorfismos como rs10919563 (*PTPRC*) (80–82), rs1801274 (*FCGR2A*) (83), rs3761847 (*TRAF1/C5*) (84), rs11591741 (*CHUK*) (17,82),

rs11541076 (*IRAK3*) (85) y rs9403 (*NFKBIB*) (17) han presentado una correlación estadísticamente significativa con desenlaces de eficacia y seguridad en pacientes con diagnóstico de AR tratados con ADA.

Teniendo en cuenta la evidencia mencionada, la AR ha demostrado ser una enfermedad cuyos pacientes podrían verse beneficiados por la aplicación de la farmacogenética, debido a que la toxicidad y efectividad de los tratamientos puede verse condicionada por variantes genéticas presentes en los pacientes y el conocimiento de estas variables en ellos puede permitir la individualización de la terapia, otorgándole un mejor perfil de seguridad y eficacia.

De esta manera, la farmacogenética puede ser utilizada como un medio de estratificación y personalización de la terapia antirreumática (86) y existe evidencia que indica que estas aproximaciones podrían ser costo-efectivas para un sistema de salud (87).

2. Modelos de aprendizaje automático en farmacogenómica

La farmacogenómica ha demostrado ser un área prometedora para la aplicación de algoritmos de aprendizaje automático, dada la gran cantidad de datos disponibles y la compleja interacción de las variables involucradas en la respuesta del paciente para algunas condiciones (88).

Por otro lado, la AR ha sido una enfermedad modelo en la farmacogenética (14,30,89–92). Sin embargo, los resultados contradictorios en estudios de asociación de único gen y la influencia simultánea de múltiples genes y condiciones predisponentes en los desenlaces de la terapia antirreumática (jugando ya sea un papel compensatorio o solapante) (7), demanda una aproximación más integral y exhaustiva en los estudios de asociación genética. De igual manera, el crecimiento exponencial en la disponibilidad de datos biomédicos demanda el uso de aproximaciones computacionales para construir modelos y tener un mejor entendimiento de la información (93). Una aproximación común para la superación de la incertidumbre entre la información encontrada en estudios epidemiológicos y su aplicación en la práctica clínica, es la utilización de algoritmos de aprendizaje automático y minería de datos para el análisis de conjuntos de datos biomédicos y el posterior desarrollo de modelos (94,95). Los desarrollos en minería de datos y aprendizaje automático en farmacogenómica han incentivado el desarrollo de estudios predictivos en esta patología (21–23,25). A continuación, se resumen algunos de los algoritmos más utilizados en farmacogenómica.

2.1. Regresión logística

La regresión logística es quizá, junto con las redes neuronales artificiales, una de las metodologías de aprendizaje automático más utilizadas en conjuntos de datos médicos, y es generalmente el estándar de comparación al momento de aplicar otras metodologías

sobre el mismo conjunto de datos (96). Este método ha sido bastante utilizado en modelos predictivos en farmacogenómica (23,25,97,98), siendo frecuentemente el primer método de prueba en proyectos que involucran diferentes algoritmos de aprendizaje automático (99–101). La regresión logística es una alternativa a la regresión lineal cuando la variable respuesta es categórica; de esta manera, la regresión logística indica la probabilidad de que, dado un conjunto inicial de predictores, la variable respuesta pertenezca a la categoría y_i, donde i puede tomar dos o más valores. Para evitar que los valores de probabilidad sean mayores que 1 o menores que 0, el conjunto de datos se ajusta a una función logística, que es de la forma:

$$p(y_i) = \frac{e^{\beta_0 + \langle X \cdot B \rangle}}{1 + e^{\beta_0 + \langle X \cdot B \rangle}}$$

donde $\langle X \cdot B \rangle$ es el producto punto entre los vectores X (vector de predictores) y B (vector de coeficientes). Esta ecuación puede transformarse de tal forma que sea lineal con respecto a los predictores y coeficientes:

$$\log\left(\frac{p(y_i)}{1 - p(y_i)}\right) = \beta_0 + \langle X \cdot B \rangle$$

El lado izquierdo de la ecuación es llamado *log-odds* o *logit*, dado que el término $\frac{p(y_i)}{1-p(y_i)}$ se denomina *odds* (o momios). Así, se puede observar que cada cambio unitario en X se traduce en un cambio unitario en los *odds* y no en la probabilidad *per se* de ser asignado a una categoría (102).

Para la estimación de cada coeficiente en B, se ajusta cada $\beta_i = \{\beta_0, \beta_1, \beta_2, \cdots, \beta_n\}$ de tal forma que la *función de verosimilitud* se maximice. Esta maximización produce que la probabilidad estimada $\hat{p}(y_i)$ de pertenecer a la categoría y_i corresponda lo más fielmente posible a la categoría real de la variable respuesta. Una visión intuitiva del proceso sería:

$$\max L(\beta_i)\,, L(\beta_i) = p(y = y_i|\beta_i)$$

Finalmente, además de servir como método de clasificación, la regresión logística también puede utilizarse como herramienta de selección de variables, debido a que tanto la

magnitud como la significación estadística de cada β_i puede utilizarse como criterio de importancia de la variable x_i como predictor del desenlace (102). Mas aún, diversos estudios predictivos en farmacogenómica han aplicado la regresión logística como algoritmo inicial para la selección de variables (103–107).

2.2. Árboles de decisión

Los árboles de decisión, también llamados árboles de regresión y clasificación (CART por sus siglas en inglés – *Classification and Regression Trees*) son métodos de aprendizaje automático utilizados por su simplicidad e interpretabilidad (102). Como su nombre lo indica, son métodos que pueden ser usados para el modelamiento de desenlaces cuantitativos (regresión) y cualitativos (clasificación) e involucran la estratificación y segmentación del espacio de predictores en regiones simples y definidas que pueden ser resumidas en forma de árbol. Sin embargo, a pesar de su interpretabilidad presentan una alta varianza, que es la tendencia de un algoritmo a generar modelos que se ajustan demasiado al ruido de fondo del conjunto de predictores e impiden la generalización (108). Este ajuste cercano al ruido de fondo también es llamado sobreajuste.

Algunos estudios predictivos en farmacogenómica han utilizado este método como punto de partida para análisis predictivos en cuanto a desenlaces terapéuticos obteniendo distintos resultados, relacionados principalmente con las características de los datos de partida (104,109,110). En general, el algoritmo de construcción de un árbol de regresión o de clasificación se compone de dos pasos (102,111):

1. Dividir el espacio de predictores (X_1, X_2, $X_3 \ldots X_n$) en J regiones distintas y mutuamente excluyentes (R_1, R_2, $R_3 \ldots R_J$), utilizando un criterio definido para su construcción. En el caso de los árboles de regresión este criterio es usualmente la minimización del residuo de la suma de cuadrados (RSS) definido por:

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j} \right)^2$$

donde $\hat{y}_{R_j}$ es el promedio de los valores de desenlace (o respuesta) para las observaciones de entrenamiento dentro del $j$avo segmento. En el caso de los árboles de clasificación, existen diferentes criterios para particionar el espacio de

predictores como el índice de Gini, la entropía o la tasa de error de clasificación. En general, el índice de Gini, también conocido como pureza de nodo, es el más usado como criterio de clasificación (102) y está definido como:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

donde $\hat{p}_{mk}$ es la proporción de observaciones de entrenamiento en la $m$ésima región que pertenecen a la $k$ésima clase. Finalmente, dado que es computacionalmente inviable considerar cada partición de los n predictores en *J* regiones, este algoritmo realiza una aproximación llamada *división binaria recursiva* que comienza en la parte superior del árbol (donde todas las observaciones pertenecen a una misma región) y divide sucesivamente el espacio de predictores en dos ramas hacia abajo del árbol. En cada paso el algoritmo realiza la mejor división posible en lugar de considerar divisiones que conduzcan a un mejor árbol en un paso futuro. Debido a esto, se dice que este algoritmo es voraz (o *greedy* en inglés) (102).

2. Posterior a la división del espacio de predictores se realiza la predicción para cada región R_j, que consiste en el promedio del valor de la respuesta para cada observación dentro de R_j para las respuestas cuantitativas (regresión) o la categoría más frecuente dentro de la región R_j para las respuestas cualitativas (clasificación) (102,111).

Los CART sufren de alta varianza, lo que implica que su resultado depende mucho de los componentes del conjunto de entrenamiento, lo que les da un pobre desempeño en comparación con otros algoritmos. Sin embargo, aunque los CART no presentan el mismo nivel de exactitud en la predicción que otros algoritmos, su desempeño predictivo puede verse significativamente mejorado por medio del uso de métodos de aprendizaje en conjunto de árboles agregados como el *boosting* y *bagging* y los bosques aleatorios (111,112).

2.2.1. Boosting

El término *boosting* (o potenciación) se refiere a una familia de métodos que mejoran el desempeño de algoritmos convirtiendo clasificadores débiles en clasificadores fuertes. El *boosting* es una aproximación general que puede aplicarse en diversos métodos de aprendizaje automático (113). Consiste en la creación secuencial de clasificadores, lo que significa que, en el caso de los CART, cada árbol es construido a partir de la información obtenida de los árboles anteriores. Sin embargo, en el *boosting* no se busca ajustar un modelo que abarque todo el conjunto de datos (por ejemplo, un árbol que tenga en cuenta todos los *n* predictores o las *i* observaciones), en cambio, se realiza un proceso de aprendizaje lento, que implica la creación de clasificadores débiles (*stumps* en el caso de los CART) cuyos errores (en clasificación) o residuos (en regresión) se tendrán en cuenta para la construcción de árboles posteriores.

Así, para los árboles de regresión, se ajustan *B* clasificadores de forma sucesiva de tal manera que para cada *b = 1,2, …, B* (102):

1. Ajustar un árbol $f_b(x)$ al conjunto de entrenamiento *X* y calcular los residuos $r_i = y_i - f_b(x_i)$
2. Actualizar el árbol $\hat{f}(x) = f(x) + \lambda f_b(x)$
3. Actualizar los residuos $r_i = r_i - \lambda f_b(x_i)$
4. Desplegar el modelo final: $\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}_b(x)$

donde el parámetro λ define cuán “rápido” o cuán “lento” aprenderá el modelo. Se puede observar que, entre menor el parámetro, menor preponderancia tendrá determinado árbol en el modelo final. En general, los métodos de aprendizaje automático que aprenden lentamente se desempeñan mejor (102)

En cuanto a los árboles de clasificación, el error de clasificación es el factor ponderador para la construcción de los árboles. Así, para un conjunto de observaciones ($\mathbf{X}_1$, y_1), ($\mathbf{X}_2$, y_2), …, ($\mathbf{X}_n$, y_n), donde $\mathbf{X}_i$ es un vector de predictores para la observación *i* y y_i es el desenlace o respuesta dicotómica de esa observación, el proceso de *boosting* consiste en (111):

1. Establecer un ponderador para cada observación ($\mathbf{X}_i$, y_i) de la primera iteración: w_{1i}. En este punto, todas las observaciones tienen la misma ponderación 1/*n*.

2. Generar un clasificador f_1 y calcular el error e_{1i} asociado al clasificador f_1 para cada *i*ésima observación:
   a. 0 si $f_1(\boldsymbol{X}_1) = y_1$
   b. 1 si $f_1(\boldsymbol{X}_1) \neq y_1$
3. La tasa de error **err**$_1$ para el clasificador f_1 se calcula entonces como:

$$\boldsymbol{err}_1 = \sum_{i=1}^{n} w_{1i} e_{1i}$$

   Lo que implica que **err**$_1$ es la proporción de observaciones que son erróneamente clasificadas por f_1

4. Se calculan nuevos ponderadores basados en el parámetro c_1, cuyo valor absoluto aumenta a medida que el error se acerca a 0,5; es decir, al poder de clasificación equivalente a lanzar una moneda.

$$c_1 = \ln \frac{1 - err_1}{err_1}$$

5. Los nuevos ponderadores son calculados de la siguiente forma:

$$w_{21} = w_{11} e^{c_1 e_{11}}$$

$$\dots$$

$$w_{2n} = w_{1n} e^{c_1 e_{1n}}$$

   Luego son normalizados, con el fin de que la suma de todos sea uno. De esta forma, es posible observar que las clasificaciones erradas tienen mayor ponderación que las clasificaciones correctas. Esta ponderación mayor hace que estas observaciones tengan mayor preponderancia en la construcción de los árboles subsiguientes.

6. La lista de ponderadores w_{21}, w_{22}, w_{22}, …, w_{2i} para cada observación ($\mathbf{X}_1$, y_1), ($\mathbf{X}_2$, y_2), …, ($\mathbf{X}_n$, y_n) se toma como una distribución de probabilidad para la construcción de un conjunto de datos (*$\mathbf{X}_1$, *y_1), (*$\mathbf{X}_2$, *y_2), …, (*$\mathbf{X}_n$, *y_n) con reemplazo. Esto

significa que aquellas observaciones con mayor ponderación estarán repetidas más veces en el nuevo conjunto de datos.

7. Finalmente, un nuevo clasificador f_2 es construido con el nuevo conjunto de datos, se computan y ponderan los errores de clasificación para cada observación y se construye un nuevo conjunto de datos.

8. Este proceso se repite *m* veces hasta que se construye un clasificador aditivo final:

$$F(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(\boldsymbol{X})$$

Donde $f_i(\boldsymbol{X})$ toma valores de 1 o 0 dependiendo del grupo en el que clasifique a la observación. Así, cada observación **X** es clasificada de acuerdo con una votación de mayorías (notar que F(x) será el porcentaje de clasificadores que asignaron una observación particular al grupo 1).

2.2.2. *Bagging*

El término *bagging* se refiere a un método de mejoramiento de desempeño general para algoritmos de aprendizaje automático y su propósito es disminuir la varianza de un algoritmo dado (102). El término *bagging* viene de *bootstrap aggregating*, debido a que, con el fin de reducir la varianza del modelo final, se realiza el ajuste de clasificadores intermedios en numerosos "nuevos" conjuntos de datos. Así, a partir de un conjunto inicial Z_0 de *n* observaciones, se crean *p* conjuntos de *n* observaciones por medio del muestreo aleatorio de observaciones con reemplazo. A este proceso se denomina *bootstraping*.

Posteriormente, para cada Z_1, Z_2, Z_3, … Z_p, se ajustan *p* modelos $\hat{f}_i(Z_i)$ de tal manera que, para los árboles de regresión, la clasificación final está dada por:

$$f_{bagg}(Z_0) = \frac{1}{p}\sum_{i=1}^{p} \hat{f}_i(Z_i)$$

Para los árboles de clasificación, el resultado del *bagging* es un voto de mayorías, donde la clase más común arrojada por todos los árboles es aquella que es asignada finalmente por el algoritmo (111).

Una forma expedita de calcular el error de clasificación de esta técnica es por medio de la estimación del error *out-of-bag* (OOB). En general, se sabe que nuevo conjunto resultado del *bagging* contiene aproximadamente un tercio de las observaciones del conjunto de datos original; las observaciones que no están incluidas en este conjunto son llamadas observaciones OOB (es decir, fuera de la bolsa) (102). De esta manera, se puede predecir la respuesta de la observación *j* utilizando los Z_i en donde la observación *j* no está presente, es decir donde esta observación es OOB. Esto producirá alrededor de p/3 predicciones que pueden ser promediadas (regresión) o cuya clasificación más frecuente puede ser extraída (clasificación). El error cuadrático medio (regresión) o error de clasificación (clasificación) para las *p* observaciones es conocido como el error OOB (102,114).

2.3. Bosques aleatorios

Los bosques aleatorios podrían verse como un caso particular del *bagging* aplicado específicamente a los árboles de decisión. Mientras que el *bagging* es un método para mejorar el desempeño de distintos algoritmos, los bosques aleatorios sólo funcionan teniendo a árboles de decisión como base (111). La principal diferencia entre los bosques aleatorios y el *bagging* es que el primero limita la correlación de los árboles que lo componen. Así, en lugar de considerar los *m* predictores de un conjunto de datos para realizar la primera partición de un árbol, cada árbol dentro del bosque considera un subconjunto de *p* predictores que usualmente equivale a $\sqrt{m}$. Esto se hace para disminuir la preponderancia de un predictor específico en la construcción de los árboles: por ejemplo, si un predictor está fuertemente relacionado con la respuesta, este va a ser siempre el primer nodo de todos los árboles construidos (dado que el algoritmo de generación de árboles de decisión es voraz o *greedy* y, por lo tanto, considera la mejor partición como primera opción), lo que hace que todos los árboles estén correlacionados entre sí. De esta manera, en promedio, un $\frac{m-p}{m}\%$ de las particiones no van a considerar aquel predictor fuertemente correlacionado con el desenlace y por ende, la varianza total del conjunto va a disminuir (102).

Si bien los bosques aleatorios nacieron como algoritmos de clasificación y regresión (115), una de sus características permite que sean utilizados también como algoritmos para la selección de variables: la estimación de la importancia de los predictores (116). Esta puede

ser expresada como la disminución total de la RSS (para árboles de regresión) o el índice de Gini (para los árboles de clasificación) en todos los árboles de los que se compone el bosque. La disminución en cada uno de estos parámetros se da debido a la partición alrededor de un predictor definido. Así, la disminución en la RSS o índice de Gini para un predictor dado se suma a través de todos los árboles por lo que, entre mayor la disminución, mayor importancia tiene la variable en el conjunto (102,116).

2.4. Máquinas de soporte vectorial

Las máquinas de soporte vectorial (SVM por sus siglas en inglés *support vector machines*) son un algoritmo de clasificación que desarrollado en la década de 1990 y que ha sido frecuentemente utilizado desde entonces para diversos problemas de clasificación (117)**.** En farmacogenómica, estos algoritmos han sido utilizados ampliamente para la predicción de desenlaces a partir de variables ómicas y variables clínicas (106,118–120), con un desempeño relativamente bueno comparado con otros clasificadores en los mismos conjuntos de datos.

2.4.1. Clasificadores de margen máximo

Las SVM son un método de clasificación de dos clases basado en encontrar el margen máximo de separación lineal entre ellas (conocido como MMC por sus siglas en inglés *maximal margin classifier*), también llamado clasificador de soporte vectorial. El MMC se basa en la generación de un hiperplano de en un espacio de *p* dimensiones (que se refiere a los *p* predictores en un conjunto de datos) para separar un conjunto de *n* puntos (*n* observaciones) maximizando la margen que hay entre los dos puntos más cercanos al hiperplano, como se explica a continuación.

En un conjunto de datos, $X_1, X_2, \cdots, X_n$ son vectores con *p* elementos (predictores) y $y_1, y_2, \cdots, y_n \in \{-1,1\}$ son las variables respuesta para cada una de las *n* observaciones, donde -1 representa una clase y 1 representa otra clase. Finalmente, el hiperplano descrito anteriormente separará las observaciones de acuerdo con la clase y_i a la que pertenezcan, de tal forma que el hiperplano para la observación i descrito por la ecuación:

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} > 0 \text{ si } y_i = 1$$

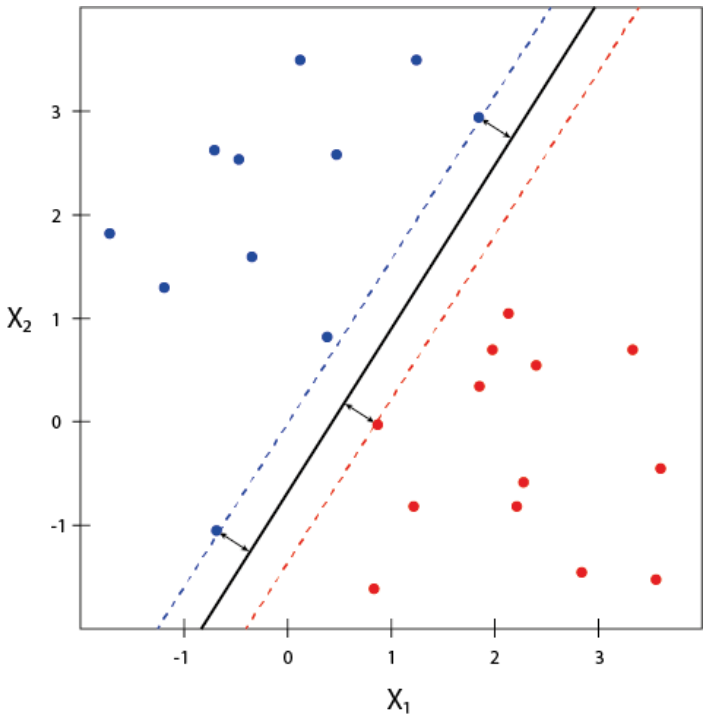
$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} < 0 \text{ si } y_i = -1$$

$$\therefore$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) > 0 \; \forall \; i \in \{1, \cdots, n\}$$

β_i son los coeficientes de la ecuación y x_{ij} se refiere al componente j del vector de predictores X_i. Esto implica que lo que esté encima o debajo del hiperplano será clasificado respectivamente como 1 y -1, como se muestra en la **Figura 2**. En esta, los puntos de color rojo corresponderían a aquellos vectores donde el resultado de la ecuación anterior sería menor que 0 y los de color azul a aquellos donde el resultado de la ecuación es mayor que 0 (102). Así, para una observación nueva X_i^*, el valor de la ecuación $\beta_0 + \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + \cdots + \beta_p x_{ip}^*$ determinará si es asignada a la clase 1 o -1.

Figura 2. Representación del hiperplano en el espacio de p dimensiones como una línea en un espacio de 2 dimensiones. La margen del hiperplano se muestra como una línea punteada



En el hiperplano de la **Figura 2**, la distancia entre el punto más cercano al hiperplano y el hiperplano es llamada la margen del hiperplano. El concepto de margen ayuda a escoger el mejor hiperplano para separar los datos, debido a que si un conjunto de datos es separable por un hiperplano, entonces existen infinitos hiperplanos que pueden separarlo (102). Así, el hiperplano de máxima margen es aquel que separa los puntos de tal forma que el valor de la margen *M* sea máximo. El clasificador que usa este hiperplano es llamado clasificador de margen máximo. Así, la ecuación del hiperplano de arriba se puede generalizar de la forma:

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M \; \forall \; i \; \in \{1, \cdots, n\}$$

2.4.2. Clasificadores de soporte vectorial

Existen problemas donde no es posible generar un hiperplano que separe estrictamente al conjunto de puntos de acuerdo con sus categorías; es decir, la ecuación $y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})$ no tiene solución con valores de $M \geq 0$. Por lo tanto, para superar esa dificultad, se han creado clasificadores que permitan la clasificación errónea de algunos puntos (localizarlos al lado incorrecto del hiperplano) con el fin de clasificar mejor la mayoría de las observaciones y ganar robustez frente a variaciones pequeñas en observaciones individuales; en otras palabras, disminuir la varianza del clasificador (121). Este tipo de clasificadores se denominan clasificadores de margen suave o clasificadores de soporte vectorial y permiten que algunos puntos queden localizados en el lado incorrecto de la margen o del hiperplano. De esta manera, para los clasificadores de soporte vectorial, el problema de maximización de *M* se modifica de la siguiente manera:

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M(1 - \varepsilon_i) \; \forall \; i \; \in \{1, \cdots, n\},$$

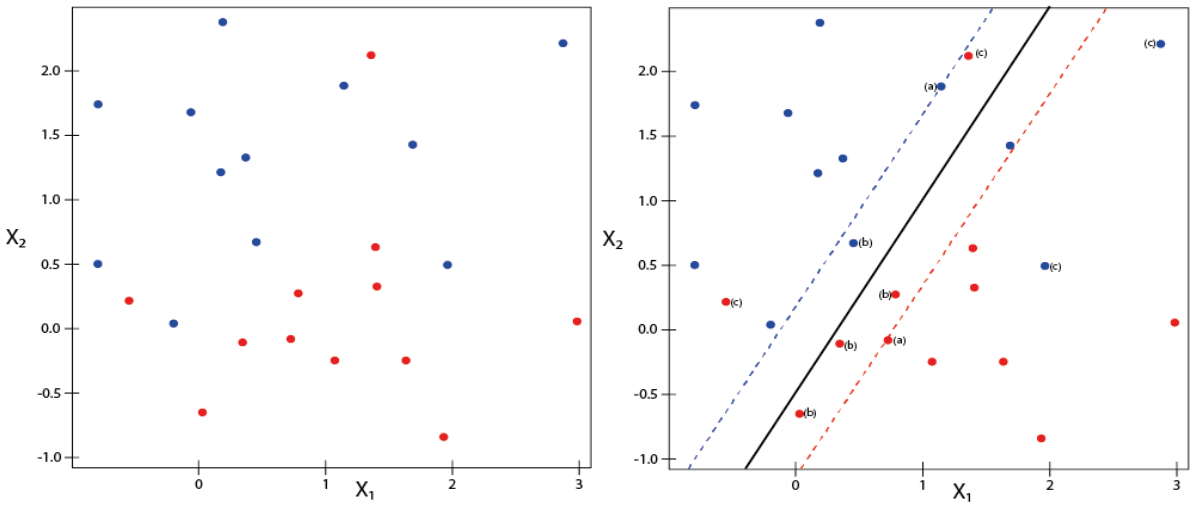
$$\varepsilon_i \geq 0, \sum_{i=1}^{n} \varepsilon_i \leq C$$

donde ε_i es un término de "error" que indica la localización del punto i con respecto a la margen: Si $\varepsilon_i = 0$, entonces i está en el lado correcto de la margen, si $0 < \varepsilon_i < 1$, entonces i está en el lado incorrecto de la margen y si $\varepsilon_i > 1$, entonces i está en el lado

incorrecto del hiperplano. C es un valor no negativo de ajuste, que limita los valores que puede tomar ε_i. Por ejemplo, si $C = 0$, entonces todos los $\varepsilon_{1,\cdots,n} = 0$, mientras que si $C > 0$, significa que no más de C observaciones pueden estar en el lado incorrecto del hiperplano (notar que, si una observación está fuera del hiperplano, entonces $\varepsilon_i > 1$ y el clasificador requiere que $\sum_{i=1}^{n} \varepsilon_i \leq C$) (102).

Además, para estos clasificadores vale la pena resaltar el fenómeno de que no todos los puntos del conjunto de datos afectan el posicionamiento del hiperplano y creación del margen: en realidad, la posición del hiperplano depende únicamente de los puntos que están en la margen, violan la margen o violan el hiperplano. Como se puede observar en la **Figura 3**, si se modifica la posición de alguno de los puntos que no están en las posiciones mencionadas a otra posición que tampoco se encuentra dentro de las mencionadas, la posición del hiperplano no se va a ver afectada. Es por esto que los puntos que se encuentran en las posiciones mencionadas son denominados vectores de soporte, porque determinan la localización del clasificador (121).

Figura 3. Representación de un conjunto de puntos que no pueden separarse perfectamente por un hiperplano. Izquierda: Distribución de los puntos. Derecha: Clasificador de soporte vectorial que permite la localización errónea de algunos puntos con respecto al hiperplano. De esta manera, se observan puntos que están en el lado correcto de la margen y del hiperplano, (a) en la margen del hiperplano, (b) en el lado correcto del hiperplano, pero en el lado incorrecto de la margen y (c) en el lado incorrecto del hiperplano.



2.4.3. Máquinas de soporte vectorial

Aunque los clasificadores de soporte vectorial permiten manejar mejor ciertos conjuntos de datos no separables perfectamente, siguen teniendo la desventaja de necesitar una frontera *lineal* para separar dos conjuntos de puntos. De esta manera, en problemas como el de la **Figura 3** los clasificadores lineales son de poca o nula utilidad, dado que no son *linealmente separables*. Para resolver este tipo de problemas se idearon las máquinas de soporte vectorial, generalizan los clasificadores lineales a espacios más grandes que el espacio de características (117,121).

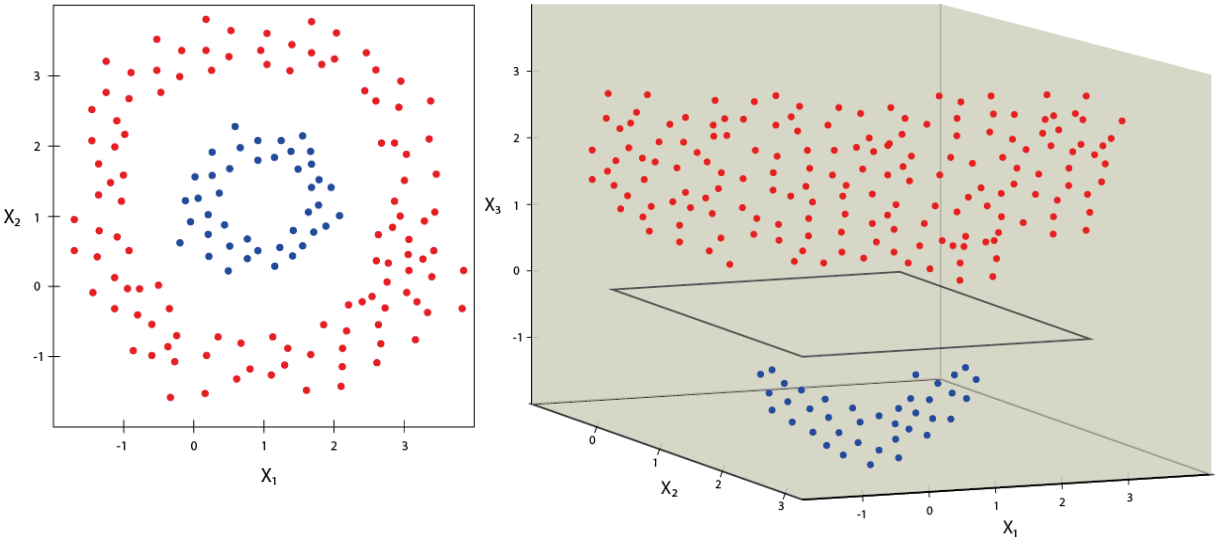
Las SVM ajustan un clasificador a un espacio de $q, q > p$ dimensiones en lugar de uno de p dimensiones utilizando funciones *kernel*. El objetivo de estas funciones es permitir la aplicación de un clasificador lineal para resolver un problema no lineal (122). De esta forma, el problema del clasificador de soporte vectorial podría convertirse en uno de la forma:

$$\max M\,, y_i\left(\beta_0 + \sum_{j=1}^{p}\beta_{j1}x_{ij} + \sum_{j=1}^{p}\beta_{j2}x_{ij}^2 + \cdots + \sum_{j=1}^{p}\beta_{ja}x_{ij}^a\right) \geq M(1-\varepsilon_i)\ \forall\ i \in \{1,\cdots,n\},$$

$$\varepsilon_i \geq 0, \sum_{i=1}^{n}\varepsilon_i \leq C$$

si se utiliza una función *kernel* polinómica. Sin embargo, existen muchas otras funciones *kernel* para mapear un vector de *n* dimensiones a un espacio de *m* dimensiones. Utilizando estas funciones, se pueden generar espacios y clasificadores como los de la **Figura 4** para separar conjuntos no linealmente separables con funciones lineales.

Figura 4. Representación de un problema no separable linealmente. Se observa la generación de una dimensión adicional para permitir la separación lineal de un problema que inicialmente no podría ser separable de forma lineal.



2.5. Redes neuronales artificiales

Las redes neuronales artificiales (ANN por sus siglas en inglés – *Artificial Neural Networks*) son quizá el algoritmo de más reciente aplicación en ciencias de la vida, así como en muchos otros campos, aunque su desarrollo data de 1943 por McCulloch y Pitts (123). En general, como su nombre lo indica, este algoritmo busca replicar el proceso de transmisión de información de las neuronas, por lo que los datos de entrada son pasados por una o múltiples capas de “neuronas”, que internamente realizan operaciones que desembocan en una salida al final del proceso (96). En farmacogenómica, las ANN han sido recientemente utilizadas dentro de los algoritmos predictivos de respuesta, debido a su poder predictivo y baja varianza (110,124–126).

Si bien el desarrollo en el campo de las ANN ha sido acelerado en los últimos años (dado principalmente al aumento del volumen de información y capacidad de cómputo) y, por tanto, existen muchas variantes de estos modelos (127–129), los perceptrones multicapa (PM) aún se encuentran dentro de las variantes más utilizadas de las ANN para el análisis de datos en salud (130–133). Los bloques fundamentales de los perceptrones multicapa

son las neuronas, por donde los componentes del conjunto de datos (los valores de entrada) son sometidos a la siguiente transformación:

$$y = f(\mathbf{w}^T\mathbf{x}) = f\left(\sum_{i=1}^{n} w_i x_i\right)$$

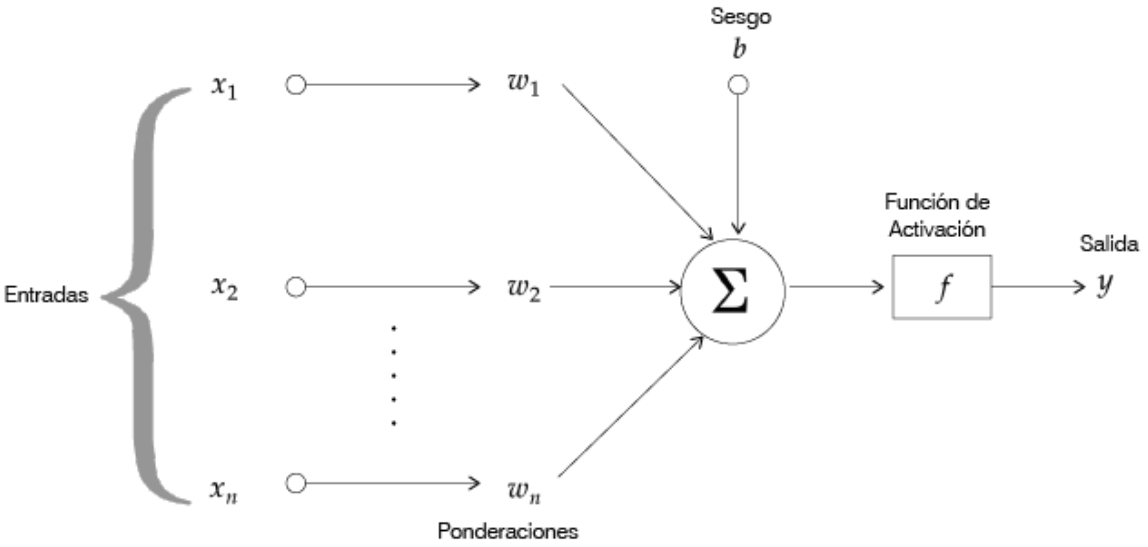
donde y es el valor de salida de la neurona, $\mathbf{x} \in \mathbb{R}^n$ es el vector de valores de entrada para una observación específica del conjunto de datos y $\mathbf{w} \in \mathbb{R}^n$ es el vector de ponderadores o pesos asignados a cada variable o característica del conjunto de valores de entrada, por lo que tanto $\mathbf{x}$ como $\mathbf{w}$ tienen que ser del mismo tamaño (n). El vector $\mathbf{w}$ cuenta con un parámetro adicional al valor de los ponderadores para cada variable: el sesgo b, que es un parámetro propio de cada neurona y por eso se encuentra en $\mathbf{w}$ y su valor correspondiente en $\mathbf{x}$ es 1. La función de este parámetro es servir como un valor umbral para determinar la activación de la neurona, como se explica a continuación.

El valor obtenido de la operación $\mathbf{w}^T\mathbf{x}$ (un escalar) es luego la entrada de la función f, llamada también función de activación y, como su nombre lo indica, determina si la neurona se activa con base en el valor del escalar. La salida de la función de activación es un valor entre 0 y 1, donde los valores cercanos a 1 indican que la neurona está “más activada”. Como se puede observar, la activación de la neurona depende tanto de los pesos como de su sesgo b, por lo tanto, se puede considerar que la activación de una neurona es decidida por el “sistema” con base en las características del conjunto de datos que cada una de ellas considera como importantes (aquellas con mayor ponderación) (96). Las funciones de activación pueden ser de diferentes tipos, las más populares son la función sigmoide (I), la tangente hiperbólica (II) y la función de activación lineal rectificada – ReLU (III).

$$\text{I: } f(x) = \frac{1}{1+e^{-x}}\ ,\ \text{II: } f(x) = \tanh(x)\ ,\ \text{III: } f(x) = \begin{cases} 0, si\ x \leq 0 \\ x, si\ x > 0 \end{cases}$$

El proceso descrito anteriormente sucede en las m neuronas que forman la capa i del perceptrón, por lo que y se convierte en el vector $\mathbf{y} = f(\mathbf{W}^T\mathbf{x}), \mathbf{W} \in \mathbb{R}^{n\times m}$ para cada capa. Las salidas de las neuronas de una capa son las entradas de las neuronas de la siguiente capa hasta llegar a la salida, como lo muestra la **Figura 5**.

Figura 5. Esquema de funcionamiento de una neurona dentro de una red neuronal artificial



El entrenamiento de este sistema se da por medio de la propagación reversa (*backpropagation* en inglés), que consiste en la definición inicial de una función de costo:

$$C(\mathbf{y}, \mathbf{o}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - o_i)^2$$

También llamada error cuadrático medio, donde es el y_i componente i del vector $\mathbf{y}$, que corresponde a la salida de la última capa de neuronas (es decir, la predicción de la red); $\mathbf{o}$ es el vector de los valores reales de la variable respuesta y o_i es el componente i del vector $\mathbf{o}$.

La matriz $\mathbf{W}$ es entonces inicializada con valores aleatorios, luego se pasa por la red la primera observación del conjunto de entrenamiento y se compara la predicción con el valor real por medio de la función de costo. Entre más lejos se encuentre la predicción del valor real, el valor de la función de costo será mayor, por lo que el objetivo del entrenamiento es reducir este valor a medida que se van pasando las diferentes observaciones del conjunto de entrenamiento (134). La minimización de esta función se realiza encontrando su gradiente ∇, que puede interpretarse como la dirección y la magnitud en la cual la función

varía más rápidamente y está definido como el vector de las derivadas parciales de la función:

$$\nabla C(\mathbf{y}, \mathbf{o}) = \begin{bmatrix} \frac{\partial C(\mathbf{y}, \mathbf{o})}{\partial \mathbf{y}} \\ \frac{\partial C(\mathbf{y}, \mathbf{o})}{\partial \mathbf{o}} \end{bmatrix}$$

Dado que el proceso de entrenamiento se resume en encontrar los valores de ponderadores que componen la matriz $\mathbf{W}$ que minimicen los valores de C y dado que $\mathbf{y}$ es una función de la matriz $\mathbf{W}$, problema de minimización se convierte en la determinación del gradiente de C con respecto a $[\mathbf{W}, \mathbf{x}, \mathbf{o}]$. El proceso de entrenamiento acaba cuando se encuentra el valor mínimo de los ponderadores, dado un conjunto de observaciones de entrenamiento y variables respuesta (135).

3. Métodos de selección de variables

Antes de entrenar un algoritmo para predecir desenlaces con base en un conjunto de predictores, es importante seleccionar las variables que deben ser utilizadas de este conjunto para realizar la predicción. De manera intuitiva, se deberían seleccionar aquellas variables que tienen mayor relación con el desenlace por predecir, es decir, las variables más relevantes. Esto se vuelve aún más crítico en conjuntos de datos de alta dimensionalidad, donde pueden existir más variables que observaciones.

La selección de variables, como su nombre lo indica, es el proceso de obtención de un subconjunto de variables a partir del conjunto inicial, de acuerdo con un criterio de selección definido. Esto permite remover variables irrelevantes, redundantes o que generen ruido en el conjunto de datos, lo que acelera el tiempo de cómputo de los algoritmos, mejora su poder predictivo y aumenta su comprensión (136). Idealmente, un proceso eficiente de selección de variables debería conservar aquellas más *relevantes* y eliminar aquellas *redundantes*. Una variable es estadísticamente relevante cuando su remoción del conjunto de datos disminuye el poder predictivo de un algoritmo; en contraste, una variable es redundante cuando su remoción no disminuye significativamente el poder predictivo en un algoritmo (137). No obstante, la plausibilidad y sentido biológico de las variables seleccionadas también debe evaluarse en estos procesos de selección, dado que un predictor puede estar estadísticamente relacionado, pero ser clínicamente irrelevante.

La selección de variables se ha utilizado en diversos estudios de aprendizaje automático en farmacogenómica (103–106), en donde muchos algoritmos diseñados originalmente para la predicción de desenlaces son utilizados para seleccionar variables, dada la posibilidad de evaluación de poder predictivo dentro del mismo algoritmo (21,100,138,139), sin embargo, métodos estadísticos paramétricos como las pruebas T y Chi-cuadrado también han sido utilizadas como método de selección de variables (101,125).

En general, cuando se trata de realizar un proceso de selección de variables no existe usualmente un camino definido; por lo tanto, se deben realizar comparaciones entre un espectro de diferentes algoritmos de selección de variables existentes para observar el panorama completo del rendimiento de los nuevos algoritmos o de los existentes (137). De esta manera, en el presente trabajo se exploraron diferentes algoritmos de selección de características, con el fin de explorar la estructura del conjunto de datos y las variables que estarían más relacionadas con el desenlace. Se realizó la exploración por separado de las variables clínicas y los polimorfismos genéticos, debido a que estos últimos fueron seleccionados *ex ante* con base en la evidencia clínica que los vinculaba al desenlace terapéutico con FARMEs.

3.1. Máxima relevancia y mínima redundancia (mRMR)

Este algoritmo fue propuesto por Peng et al., (2005) (140) y está basado en el criterio de información mutua para seleccionar las variables más relevantes (máxima relevancia). De esta manera, para dos variables aleatorias continuas x y y, su información mutua, definida en términos de sus funciones de densidad de probabilidad $p(x), p(y), p(x,y)$ es:

$$I(x,y) = \iint p(x,y)\ \log\frac{p(x,y)}{p(x)p(y)}dxdy$$

Para variables discretas sería:

$$I(x,y) = \sum_{i,j}^{a,b} p(x_i,y_i)\ \log\frac{p(x_i,y_i)}{p(x_i)p(y_i)}$$

Así, las variables x_i con máxima relevancia son aquellas que presentan la mayor información mutua con el predictor c: $I(x_i,c)$, lo que refleja una dependencia significativa del predictor con el desenlace. Así, un algoritmo de selección secuencial de variables extraería las m variables con mayor valor de I. Sin embargo, la combinación de variables individuales con alta correlación con el desenlace no implica necesariamente un poder de clasificación alto (141,142), esto puede deberse a que hay variables codependientes cuyo poder discriminatorio no cambia si una de ellas es retirada, por lo que la inclusión de

muchas de estas puede generar ruido al aplicar el algoritmo (143). Entonces, dado un subconjunto de variables seleccionadas S, la condición de máxima relevancia se da cuando:

$$\max D(S, c), \qquad D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i, c)$$

Y la mínima redundancia:

$$\min R(S), \qquad R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j)$$

El criterio que combina los dos parámetros anteriores es llamado “máxima relevancia y mínima redundancia” (mRMR). En la práctica, los algoritmos implementados para seleccionar variables basados en este concepto realizan una búsqueda incremental para encontrar los x_i y x_j que minimicen el valor de R y los x_i que maximicen el valor de D (140).

3.2. Análisis de correspondencias múltiples

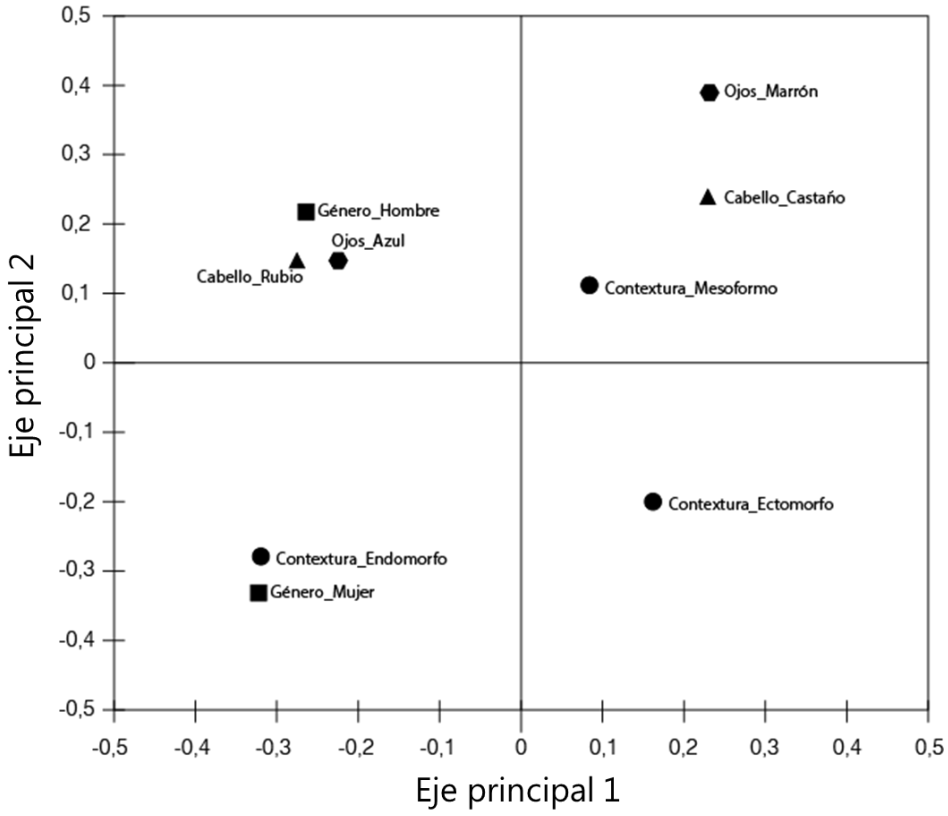
El análisis de correspondencias múltiples (ACM) es una extensión del análisis de correspondencias simple, una técnica de análisis exploratorio de datos categóricos en matrices de confusión de 2x2 (144). Como su nombre lo indica, el ACM está diseñado para conjuntos de datos multivariados, aunque mantiene el proceso analítico del análisis de correspondencias simple. El ACM permite detectar asociaciones sistemáticas entre las categorías de las variables; así, en comparación con otras técnicas de análisis de datos categóricos como chi-cuadrado, el test exacto de Fisher, entre otros, el ACM no sólo revela la asociación entre dos variables, sino también *cómo* esas asociaciones están construidas y cómo variables adicionales están correlacionadas de igual manera (144). Específicamente, el ACM se utiliza para representar conjuntos de datos como nubes de puntos en un espacio euclidiano multidimensional, donde la interpretación se basa en la posición relativa de los puntos y cómo estos se distribuyen en el espacio multidimensional. De esta manera, variables con categorías distribuidas de forma similar, tenderán a estar más cerca en la nube de puntos (145,146). Para observar cómo trabaja en la práctica un

ACM, la **Tabla 3** muestra un ejemplo de variables demográficas de un número de personas y la **Figura 6** muestra el resultado del ACM.

Tabla 3. Conjunto de datos de ejemplo con variables categóricas relacionadas con características físicas de un grupo de individuos

Individuo	Género	Color de ojos	Color de cabello	Contextura
1	F	Azul	Rubio	Endomorfo
2	M	Marrón	Negro	Ectomorfo
3	M	Marrón	Castaño	Ectomorfo
…	…	…	…	…
1000	F	Verde	Rubio	Mesomorfo

Figura 6. Análisis de correspondencias múltiples del conjunto de datos de la Tabla 3



En este ejemplo se puede notar que los hombres de esta cohorte tienden a tener el cabello rubio y ojos azules, mientras que las mujeres tienden a ser endomorfas (propensión a almacenar grasa). La distribución de los mesomórficos parece ser independiente de la distribución de categorías de las otras variables.

3.3. Eliminación recursiva de características (RFE)

Como su nombre lo indica, la eliminación recursiva de características (RFE por sus siglas en inglés – *Recursive Feature Elimination*) es un algoritmo para la selección de variables que se basa en la eliminación sucesiva de predictores del conjunto de datos y su posterior evaluación de la exactitud en la clasificación (147). Por lo tanto, este algoritmo se utiliza en conjunto con otros métodos de aprendizaje automático, que se encargan del proceso de clasificación. El funcionamiento del algoritmo se resume a continuación:

1. Inicialmente, se construyen diferentes conjuntos de entrenamiento y de prueba a partir del mismo conjunto de datos (ya sea dividiéndolo en *k* conjuntos para validación cruzada o por medio del *bootstraping*)
2. Posteriormente, se utiliza un método de aprendizaje automático para ajustar un modelo al conjunto de datos con todos los predictores y se calcula la importancia de cada predictor en la construcción del modelo.
3. Se define con conjunto *S* que contiene una secuencia del número de variables que podrían retenerse, de tal forma que S: $\{S_1 > S_2 > S_3 > S_4 \cdots S_n\}$.
4. Para cada iteración, se retienen los primeros S_i predictores más relevantes del paso 2, se ajusta de nuevo el modelo únicamente con estos predictores y se calcula su exactitud con el conjunto de prueba.
5. El número de variables a retener será aquel S_i con la mayor exactitud.
6. Realizar los pasos anteriores para cada uno de los conjuntos de entrenamiento y prueba generados. En general, se mantendrá aquel S_i más frecuente en las *k* iteraciones.

Se utilizan diferentes conjuntos de entrenamiento y prueba a partir del conjunto principal de datos, con el fin de evitar el sobreajuste de los modelos internos, debido a que, si un número grande de predictores es utilizado para crear el modelo y uno de estos predictores se correlaciona aleatoriamente con el resultado, el algoritmo RFE asignaría un buen poder de clasificación a esta variable y el error de predicción (en el mismo conjunto de datos) se reduciría. Por lo tanto, se requeriría de diferentes conjuntos de prueba para darse cuenta que este predictor no es informativo (148).

Dada la amplia disponibilidad de información ómica y la utilización cada vez más frecuente de métodos de aprendizaje automático dentro de las ciencias de la vida, el presente trabajo tiene como objetivos:

3.4. Objetivo general

- Desarrollar un modelo de asociación para los desenlaces de la terapia farmacológica con MTX y ADA en pacientes diagnosticados con artritis reumatoide, a partir de variantes genéticas y parámetros clínicos de riesgo de una cohorte de pacientes del Hospital Militar Central

3.5. Objetivos específicos

- Realizar una exploración de las diferentes asociaciones entre las variables contenidas en la historia clínica y polimorfismos genéticos, con los desenlaces de la terapia con MTX y ADA.

- Identificar las variables clínicas y genéticas que tendrían mayor impacto en el desenlace de la terapia farmacológica con MTX y ADA en pacientes diagnosticados con artritis reumatoide en el Hospital Militar Central.

- Determinar los métodos de integración y algoritmos de aprendizaje automático que mayor valor predictivo tendrían sobre el conjunto de datos genéticos y clínicos.

4. Revisión de métodos para su aplicación en conjuntos de datos biomédicos

Con el fin de determinar los algoritmos de aprendizaje automático más adecuados para aplicar a los conjuntos de datos en farmacogenética, se realizó una revisión sistemática de literatura, que también sirvió para informar las tendencias en investigación en el campo, así como los principales retos y fuentes de datos en proyectos de análisis de datos de mundo real (RWD por sus siglas en inglés – *Real World Data*) en farmacogenómica.

Se realizó una revisión sistemática de la literatura, siguiendo las recomendaciones metodológicas del manual Cochrane para revisiones sistemáticas de intervenciones (149) y aquellas de los estándares PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) (150) para el reporte de los resultados de la revisión. Así mismo, se siguieron las recomendaciones de Kitchenham et al., (151) y Mariano et al., (152) para la realización de revisiones de literatura en ingeniería de software y bioinformática, respectivamente. La metodología para el desarrollo de la revisión se documenta a continuación:

Definición de las preguntas de investigación

De acuerdo con la creciente utilización de algoritmos de aprendizaje automático en diferentes campos de las ciencias de la vida, se formularon dos preguntas de investigación que guiaron el desarrollo de la revisión sistemática, enfocadas en establecer en qué medida se han aplicado estos algoritmos en la predicción de desenlaces terapéuticos utilizando información ómica como base:

- ¿Qué métodos de modelación con fines predictivos están siendo utilizados en farmacogenómica?
- ¿Qué tipo de información ómica y no ómica está siendo utilizada para el desarrollo de los modelos?

Criterios de inclusión y exclusión

Los criterios de elegibilidad de los estudios se establecieron con base en la utilización prevista de la evidencia; es decir, la elección de algoritmos de aprendizaje automático que serían sujeto de comparación en este trabajo. Así, se establecieron los siguientes criterios de inclusión:

- Estudios experimentales (también llamados investigaciones originales) en los que se utilizara información de vida real, incluyendo estudios clínicos, estudios observacionales y bases de datos de pacientes.
- Estudios en los que se utilizara al menos una fuente de información ómica (genómica, transcriptómica, proteómica).
- Estudios en los que se utilizara algún algoritmo de aprendizaje automático o cualquier otro método de modelamiento, cuyo fin fuera la predicción de los desenlaces clínicos del tratamiento farmacológico.
- Estudios en texto completo en idioma español o inglés.

El criterio de exclusión aplicado fue:

- Estudios que no detallan los algoritmos predictivos utilizados

Adicionalmente, los estudios en los que se utilizara el mismo conjunto de datos o algoritmo/modelo (en donde fuera detectable), serían considerados como duplicados se conservaría aquella referencia con la fecha de publicación más antigua.

Búsqueda en bases de datos

Se realizó una búsqueda sistemática en las siguientes bases de datos: MEDLINE/PubMed, EMBASE, Scopus y Web of Science el 17 de noviembre de 2018, con el fin de identificar publicaciones indexadas que respondieran las preguntas de investigación planteadas. Se diseñó inicialmente una estrategia de búsqueda genérica con base en los términos clave establecidos en la pregunta de investigación. Se realizaron revisiones exploratorias de estudios de asociación genética, con el fin de refinar las palabras clave y capturar los sinónimos relevantes que se incluirían en la estrategia de búsqueda. Así, en la estrategia de búsqueda genérica se incluyeron los términos “pharmacogenomics”, “genetic association”, “genetic models”, “genetic polymorphisms”, “population genetics”, “machine learning”, “statistical learning”, “statistics”, “medical informatics”, “data mining”, “knowledge

discovery in databases", "multivariate análisis", "classification methods", "drug treatment" y "treatment outcomes". Esta estrategia de búsqueda se refinó posteriormente, adicionando sinónimos, abreviaturas, acrónimos, variaciones ortográficas, plurales y vocabulario controlado (como MeSH y Emtree para PubMed y EMBASE, respectivamente) para cada base de datos. La sintaxis se complementó con expansión de términos controlados, identificadores de campo, truncadores, operadores de proximidad y operadores booleanos, y se adaptó para cada fuente de información. Con el fin de actualizar la revisión durante su ejecución, se programaron alertas automáticas de búsqueda en todas las bases de datos consultadas, utilizando la estrategia de búsqueda usada en cada una. Así, el cuerpo de evidencia se iba actualizando mensualmente durante la tamización de resultados. El punto de corte final de adición de nuevas publicaciones a la presente revisión fue el 30 de mayo de 2019.

Para efectos de esta revisión, los conceptos "aprendizaje estadístico", "aprendizaje maquinal" y "minería de datos" se consideraron igualmente relevantes para las aplicaciones incluidas, debido a que comparten el objetivo de entender la estructura y realizar predicciones sobre un conjunto de datos.

Adicionalmente, se programó el envío automático de notificaciones de nuevos resultados de búsqueda de cada estrategia por una ventana de 6 meses, con fecha de corte del 30 de mayo de 2019. Las nuevas referencias identificadas fueron adicionadas paulatinamente al conjunto de referencias inicial. El resumen detallado de las estrategias de búsqueda se encuentra en el **Anexo 1**. La administración de las referencias encontradas se realizó en el software Mendeley 1.19.4.

Selección de referencias

Se realizó la examinación de título y resumen de cada referencia, así como una lectura diagonal (152) de la publicación, con el fin de evaluar su elegibilidad para inclusión en la síntesis de evidencia. La tamización de referencias basada en título y resumen se llevó a cabo por 2 revisores en forma independiente. La metodología para la evaluación preliminar de las referencias por título y abstract, así como la posterior lectura diagonal para confirmar el cumplimiento de los criterios de inclusión y exclusión, fue tomada del estudio de Mariano et al. (152).

Las referencias seleccionadas en la sección anterior (que cumplieron con los criterios de inclusión y fueron filtradas por el criterio de exclusión) fueron posteriormente analizadas en texto completo. Esta lectura sirvió para la evaluación de calidad/pertinencia y extracción de información de cada referencia.

La evaluación de calidad/pertinencia de las referencias consistió en un análisis empírico de cuán adecuada era cada publicación para la resolución de la pregunta de investigación. Así, si bien esto constituye un análisis que depende únicamente del criterio del evaluador, se definió una metodología estandarizada de evaluación por medio de un sistema de puntuación de 1 a 10 en 5 preguntas:

1. ¿El objetivo del estudio es un análisis predictivo?
2. ¿Se describe claramente la información utilizada para la predicción?
3. ¿Se describen claramente los algoritmos utilizados?
4. ¿Se realiza una prueba con un conjunto de datos independiente?
5. ¿Se describen claramente las métricas de desempeño del algoritmo?

El objetivo de esta evaluación fue establecer (así fuese de una manera empírica) un criterio de calidad metodológica y adecuación de las referencias a las preguntas de investigación, debido a que la revisión arrojaría un amplio espectro de metodologías para las que no existiría una única lista de chequeo de calidad. Sin embargo, las preguntas de evaluación estuvieron basadas en las consideraciones mencionadas en el TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) *statement* (153). Las referencias fueron consideradas “adecuadas” para responder la pregunta de investigación si su puntaje promedio de evaluación era de 6 o más.

Extracción de información y síntesis de evidencia

De las referencias analizadas en texto completo se extrajo la siguiente información:

- Población
- Información ómica utilizada
- Predictores
- Intervenciones
- Desenlaces
- Tamaño de muestra

- Técnica de clasificación
- Breve resumen de los hallazgos
- Exactitud en la clasificación (cualquier medida considerada por los autores)
- Sensibilidad
- Especificidad

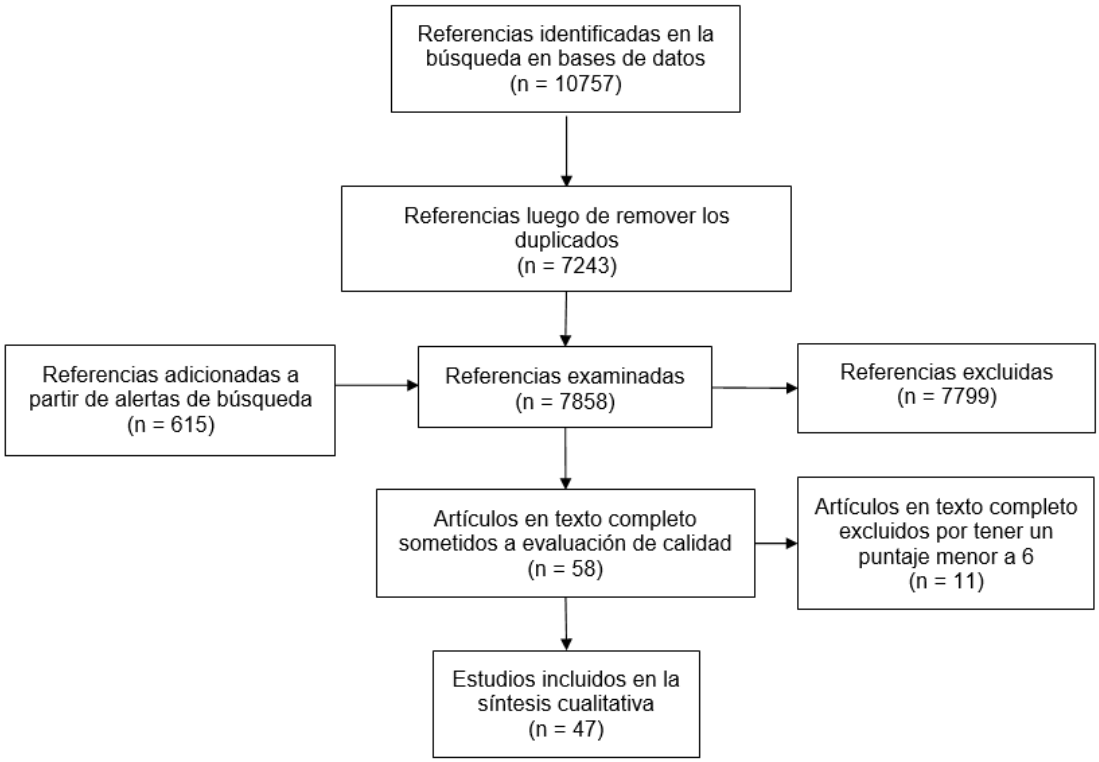
Se realizó una síntesis cualitativa de la evidencia, debido a que los estudios identificados correspondían a diferentes poblaciones y, por tanto, la eficiencia de los algoritmos predictivos no era comparable. Se escogió presentar la evidencia en forma de tablas resumen para su posterior discusión.

4.1. Resultados de la revisión

La **Figura 7** presenta el diagrama PRISMA de los resultados de la revisión. La búsqueda inicial generó un total de 7243 referencias únicas, a las que 615 referencias nuevas, provenientes de las alertas de búsqueda, fueron adicionadas. Finalmente, 58 referencias fueron revisadas en texto completo y 47 fueron incluidas en la síntesis de evidencia. La **Tabla 4** resume la puntuación que tuvieron las referencias examinadas en texto completo con respecto a las preguntas de evaluación de pertinencia/calidad metodológica. El **Anexo 2** muestra la información extraída de las referencias seleccionadas.

Se excluyeron 11 referencias de la síntesis final, debido a que tuvieron un puntaje de evaluación menor a 6 en promedio. Las razones principales de obtención de bajos puntajes fueron la falta de claridad en la descripción de los algoritmos predictivos aplicados, así como la ausencia o falta de claridad en la descripción de las medidas de desempeño de estos; que son necesarias para la comparación y análisis crítico de su utilidad en estudios predictivos.

Figura 7. Diagrama PRISMA con los resultados de la revisión



El rango de fechas de publicación va desde 2005 a 2018, siendo los últimos 5 años aquellos donde ha habido mayor densidad de publicaciones (63%). De igual manera, se observó una tendencia creciente en el número de publicaciones incluidas a medida que avanzan los años (6 publicaciones en 2015, 8 en 2016, 5 en 2017 y 10 en 2018), lo que podría indicar un crecimiento en la investigación en el área. Esto también se evidencia con el aumento en la complejidad de las estrategias, modelos y algoritmos utilizados en la predicción de desenlaces (99,154) y el creciente uso de información multi-ómica en estos estudios (155).

Tabla 4. Resultados de la evaluación de pertinencia de las publicaciones con respecto a las preguntas. Resaltadas se encuentran las publicaciones descartadas de la síntesis

Autor, año	Pregunta 1	Pregunta 2	Pregunta 3	Pregunta 4	Pregunta 5	Puntaje promedio
Rashkin et al., (2018)	10	10	10	0	8	7,6
Naushad et al., (2018)	10	10	7	0	8	7

Autor, año	Pregunta 1	Pregunta 2	Pregunta 3	Pregunta 4	Pregunta 5	Puntaje promedio
De Rotte et al., (2018)	10	10	10	10	10	10
Maciukiewicz et al., (2018)	10	10	10	0	10	8
Lin et al., (2018)	10	10	7	0	10	7,4
Zhang et al., (2018)	10	10	10	10	8	9,6
Xie et al., (2018)	10	10	8	0	10	7,6
Jenko et al., (2018)	10	10	10	10	8	9,6
Lee et al., (2018)	10	10	10	0	8	7,6
Kappel Buhl et al., (2018)	10	10	5	0	0	**5**
Lee et al., (2017)	10	3	5	0	5	**4,6**
Yang et al., (2017)	10	10	10	0	10	8
Shu et al., (2017)	10	10	0	0	0	**4**
Kim et al., (2017)	10	10	0	5	0	**5**
Kuo et al., (2017)	10	10	10	0	10	8
Jenko et al., (2016)	10	10	10	0	8	7,6
Yin et al., (2016)	10	10	8	10	10	9,6
Glass et al., (2016)	10	7	0	0	8	**5**
Kureshi et al., (2016)	10	10	8	0	5	6,6
Gonzalez Bosquet et al., (2016)	10	7	10	10	10	9,4
Shahid et al., (2016)	10	10	10	10	0	8
Liang et al., (2016)	10	5	10	0	5	6
Rizk et al., (2016)	10	10	10	0	10	8
van Dijkhuizen et al., (2015)	10	10	10	0	8	7,6
Jung et al., (2015)	7	10	7	0	0	**4,8**
Walter et al., (2015)	10	10	10	0	10	8
Kim et al., (2015)	10	10	10	0	10	8
Kautzky et al., (2015)	10	10	10	10	6	9,2
Takahashi et al., (2015)	10	10	6	0	10	7,2
Chen et al., (2014)	10	6	8	10	6	8
Kim et al., (2014)	10	8	10	10	6	8,8
van Asten et al., (2014)	10	10	10	0	10	8
KayvanJoo et al., (2014)	10	8	10	0	10	7,6
Kurosaki et al., (2013)	7	10	4	0	0	**4,2**
Beerenwinkel et al., (2013)	10	10	10	0	10	8
Zhang et al., (2013)	10	7	10	0	8	7

Autor, año	Pregunta 1	Pregunta 2	Pregunta 3	Pregunta 4	Pregunta 5	Puntaje promedio
Jiménez-Sousa et al., (2013)	6	10	10	0	5	6,2
Fransen et al., (2012)	10	10	10	0	10	8
Vidal-Castiñeira et al., (2012)	10	10	10	0	8	7,6
Neukam et al., (2012)	10	10	8	7	10	9
O'Brien et al., (2012)	10	10	10	10	10	10
Alterovitz et al., (2011)	10	10	10	0	8	7,6
Kurosaki et al., (2011)	10	10	10	10	0	8
Xu et al., (2011)	10	8	10	10	10	9,6
Caocci et al., (2010)	10	10	10	10	7	9,4
Wu et al., (2010)	10	10	5	0	0	**5**
Daemen et al., (2009)	10	7	10	0	10	7,4
Han et al., (2009)	7	10	5	0	0	**4,4**
Petrovsky et al., (2009)	10	8	10	10	10	9,6
Wu et al., (2008)	10	10	8	0	10	7,6
Lin et al., (2008)	10	10	10	10	10	10
Loi et al., (2008)	10	10	8	10	0	7,6
Hlavati et al., (2007)	10	6	10	0	0	**5,2**
Wessels et al., (2007)	10	10	10	10	10	10
Linke et al., (2006)	10	10	10	0	7	7,4
Lin et al., (2006)	10	10	10	0	10	8
Modlich et al., (2005)	10	10	7	10	5	8,4
Hakonarson et al., (2005)	5	10	0	0	6	**4,2**

Diseño de los estudios y participantes

El cáncer fue la patología mayormente estudiada en los estudios incluidos (19 estudios), seguida de la artritis reumatoide (7 estudios). De igual manera, en todos los estudios, la variable respuesta principal fue categórica, ya sea porque desde su definición es concebida así o debido a la discretización de escalas continuas. La información utilizada para la predicción procedió de dos fuentes principalmente: información observacional y estudios clínicos; en todos los casos la información fue recogida y analizada retrospectivamente. Se identificaron 2 estudios en los que se utilizaron conjuntos de datos públicos para entrenar los algoritmos (139,156). En 17 estudios la intervención fue monoterapia

farmacológica, en los demás los pacientes recibían tratamiento combinado o en esquemas, a excepción de un estudio donde la intervención era el trasplante de células precursoras hematopoyéticas (124).

Información multi-ómica

Los SNPs fueron los predictores ómicos mayormente utilizados en los estudios incluidos, específicamente en 31 publicaciones se utilizaron estos polimorfismos con o sin información clínica adicional para realizar predicciones de desenlaces terapéuticos. Por otro lado, los predictores de transcriptoma (mRNA, miRNA) fueron utilizados en 11 publicaciones, de las que el 91% (10 estudios) tenían al cáncer como patología de estudio. Se identificó una única publicación donde hubo utilización de datos de proteoma (junto con datos de transcriptoma) como predictores del desenlace (119). En los 5 estudios restantes se utilizaron las mutaciones genómicas como predictores del desenlace: En 3 estudios las mutaciones estudiadas se encontraban en el genoma de células neoplásicas (109,157,158) y en 2 estudios las mutaciones correspondían a aquellas del genoma del virus de la hepatitis C (110,159).

Además, se identificaron 6 estudios en donde se realizó genotipificación de genoma completo al inicio con el fin seleccionar las variaciones más fuertemente relacionadas con los desenlaces terapéuticos de interés; estos estudios predictivos fueron realizados en el marco de estudios de asociación de genoma completo (GWAS) (103,106,138,154,160,161). Finalmente, no se identificó ninguna publicación en la que se utilizaran las 3 fuentes ómicas principales (genómica, transcriptómica y proteómica) como predictores combinados.

Tendencias en la aplicación de algoritmos de aprendizaje automático

En 29 estudios (62%) se utilizó un único algoritmo para la predicción de los desenlaces, mientras que en los otros 20 estudios se utilizaron 2 o más algoritmos en el proceso de predicción. En general, se identificó la utilización de algoritmos para la selección inicial de variables en 24 estudios; en 18 de ellos la metodología de selección estuvo basada en regresiones y 4 aplicaron pruebas paramétricas (test Chi-cuadrado, t de student) y no paramétricas (test de Wilcoxon, Kolmogorov-Smirnov) para seleccionar variables. En el estudio de KayvanJoo et al., (110) se utilizaron también algoritmos de agrupación en

combinación con los análisis de componentes principales y máquinas de soporte vectorial para la selección de variables.

Se identificaron 6 referencias en las que el algoritmo predictivo consistió en redes neuronales artificiales. En 25 estudios se utilizaron métodos de regresión como algoritmos predictivos, 8 estudios utilizaron árboles de decisión, 7 utilizaron bosques aleatorios, 8 utilizaron máquinas de soporte vectorial, 2 utilizaron k vecinos más próximos y uno utilizó *k-means* (Tabla 2). Algunos estudios utilizaron algoritmos desarrollados *in house* o algoritmos no convencionales para la selección de variables y predicción de desenlaces, como el *Knowledge-based SNP selection* para la selección de SNPs a partir de información de GWAS (138) o el sistema de puntuación ponderado de riesgo genético que tiene en cuenta la razón de momios (OR, por su sigla *odds ratio* en inglés) univariante de cada SNP frente al desenlace (160).

La exactitud en la predicción de los algoritmos fue medida principalmente en términos del área bajo la curva (AUC) de características operativas del receptor (ROC, acrónimo de *Receiver Operating Characteristic*). Algunos estudios también reportaron la sensibilidad y especificidad para los puntos de corte establecidos luego del análisis del desempeño de cada modelo (Tabla 2). De igual manera, se identificaron estudios que compararon el desempeño de los modelos cuando se incluían variables únicamente clínicas y cuando se incluían variables genéticas y clínicas. En general, los modelos que incluyeron ambas variables se desempeñaron mejor cuando se compararon con aquellos que sólo usaron una única fuente de datos (97,162,163). En 11 estudios los predictores correspondieron únicamente a variables ómicas (SNPs, expresión diferencial de RNA y mutaciones en el genoma).

En 19 estudios se identificó la utilización de conjuntos de entrenamiento y de prueba. En estos casos, los dos conjuntos podían provenir de fuentes independientes (cohortes independientes) o podían ser fruto de una subdivisión de la cohorte disponible para el análisis. Para efectos de la puntuación de calidad también se consideraron como independientes los datos de prueba que provenían de la cohorte original, aunque se podría argumentar que estos muy probablemente comparten la misma estructura interna que la cohorte de entrenamiento.

4.2. Discusión y conclusiones de la revisión

La utilización de técnicas de aprendizaje automático para la predicción de desenlaces utilizando información ómica es un campo de estudio naciente con una tendencia creciente en los últimos años, una afirmación soportada por los hallazgos de esta revisión, en los que se evidencia que la mayoría de los estudios incluidos en la síntesis de evidencia fueron realizados en los últimos 5 años. De igual manera, no se evidenció la utilización rutinaria de este tipo de aproximaciones en la práctica clínica en ninguna publicación, debido a la dificultad de extrapolación de los hallazgos entre poblaciones diferentes (22,99,100), así como la ausencia de información ómica de pacientes en contextos diferentes al investigativo.

Con el fin de incrementar la exactitud de la predicción, remover información irrelevante y mejorar la interpretación de los resultados, la reducción de dimensionalidad es un paso de preprocesamiento de información bastante utilizado en los proyectos de minería de datos y aprendizaje automático (164). En general, existen dos aproximaciones a la reducción de la dimensionalidad: la extracción de características y la selección de características. En la primera, el objetivo es generar una transformación del espacio de las entradas, con el fin de generar un subespacio con menos dimensiones, pero que retenga toda la información relevante. Sin embargo, una de las principales desventajas de utilizar esta estrategia es que el modelo obtenido pierde su interpretabilidad (165). En contraste, la selección de características consiste en la escogencia de las variables de entrada que contengan la mayor cantidad de información relevante para la resolución del problema particular, por lo tanto, la interpretabilidad del modelo final no se ve afectada (164). Debido a la alta dimensionalidad que presentan los conjuntos de datos biomédicos y genéticos, este paso de preprocesamiento se vuelve bastante relevante en la construcción de modelos de aprendizaje automático basado en estos conjuntos de datos. Los hallazgos de esta revisión indican que la tendencia más común a la hora de construir estos modelos es su interpretabilidad, lo que se infiere por el hecho de no encontrar publicaciones cuyo paso de reducción de dimensionalidad hubiese consistido en la extracción de características. Estos hallazgos son consistentes con el objetivo final de la utilización de este tipo de datos, teniendo en cuenta que la mayoría de los esfuerzos en cuanto al análisis de esta información se enfocan en el apoyo de decisiones clínicas y terapéuticas.

Por otra parte, se ha teorizado que tanto el desenlace clínico como la farmacocinética y farmacodinamia de los fármacos son influenciados no sólo por “variables ómicas” como la genómica, transcriptómica, metabolómica, interactómica, entre otras, propias del individuo, sino también por variables epigenéticas, clínicas, nutricionales, ambientales y psicosociales (7). Así, los modelos predictivos también deberían considerar estas variables. En esta revisión, no se identificaron aproximaciones que utilizaran información nutricional o ambiental, probablemente debido a la dificultad de recolección para todos los pacientes, así como la baja probabilidad de disponibilidad de esta información en la práctica clínica real. Además, la aplicabilidad de estos modelos en la práctica clínica depende de su versatilidad y relativa simplicidad, lo que implica mayoritariamente la utilización de las variables más relevantes para el desenlace. De igual manera, se debe tener en cuenta que la finalidad de estos modelos es predictiva y no explicativa, por lo tanto, la inclusión de todas las variables estadísticamente significativas relacionadas con el desenlace no es imperativa (166).

Una de las aproximaciones más utilizadas para reducir el sobreajuste y mejorar la generalización de los resultados de los modelos es separar el conjunto de datos en conjuntos de entrenamiento, validación y prueba, con el fin de crear el modelo, optimizar sus parámetros y evaluar su desempeño, respectivamente (167). En la revisión, el 40,7% de las publicaciones utilizaron cohortes de prueba (independientes o no) para evaluar los modelos, lo que indica que la práctica aún no es tan generalizada en estudios farmacogenómicos y podría deberse a restricciones del tamaño de muestra, disponibilidad de información, entre otras, que disminuyen el poder estadístico y, por tanto, limitan la posibilidad de identificar predictores relevantes. Así mismo, es difícil establecer un *gold standard* (valor de referencia) o estimar un tamaño de muestra mínimo para el desarrollo y entrenamiento óptimo de modelos de aprendizaje automático (168). En la evidencia recopilada, los tamaños de muestra variaron entre 40 y 9231 pacientes.

Como cualquier revisión sistemática, una de las principales limitaciones de esta revisión es su susceptibilidad al sesgo de publicación. Si bien existen métodos para estimar este sesgo en metaanálisis (169), debido a la heterogeneidad de la evidencia recopilada en esta revisión, no se consideró la utilización de estimadores de sesgo de publicación en esta investigación.

Conclusiones de la revisión

La utilización de algoritmos de aprendizaje automático en farmacogenética es un campo naciente con un crecimiento elevado en los últimos años y tiene el potencial de convertirse en una poderosa herramienta de apoyo en la toma de decisiones clínicas, sin embargo, existen aún algunas barreras que superar antes de implementar estas metodologías en la práctica clínica, como lo son la estandarización de metodologías para la recolección y análisis de información clínica y ómica y la inclusión en los algoritmos de otras fuentes de heterogeneidad que podrían afectar los desenlaces terapéuticos, como las variables sociodemográficas, nutricionales y epigenéticas, entre otras. Sin embargo, los algoritmos de aprendizaje automático son herramientas analíticas poderosas para el análisis de conjuntos de datos de alta dimensionalidad, que, a diferencia de las técnicas estadísticas de comparación univariada comúnmente utilizadas en epidemiología, permiten integrar múltiples fuentes y tipos de datos para la construcción de modelos predictivos.

Los puntos de investigación futura en este campo se encuentran en la aplicación de otro tipo de algoritmos, como los algoritmos evolutivos, la exploración de diferentes arquitecturas de redes neuronales y la utilización de aproximaciones integrativas basadas en redes. En este último aspecto, la biología de sistemas puede ayudar a elucidar los mecanismos de regulación génica que dan lugar a las diferencias en efectividad y seguridad de fármacos a partir de variaciones polimórficas en determinados genes (174). De esta manera, los nuevos modelos creados a partir de información ómica y clínica no sólo serían predictivos, sino también explicativos, dando lugar a avances en el entendimiento de enfermedades que se podrían traducir rápidamente en intervenciones al paciente.

5. Metodología

El presente estudio fue realizado en el Hospital Militar Central, bajo el nombre de protocolo "Modelo farmacogenético y clínico para la predicción de desenlaces en pacientes con artritis reumatoide tratados con metotrexato y adalimumab", código 2018-087. Los pacientes participantes accedieron a compartir su información clínica y muestras de sangre para la elaboración de proyectos de investigación en farmacogenética. La firma de consentimientos informados, así como la recolección de las muestras de sangre se llevaron a cabo en el proyecto "ASOCIACIÓN ENTRE LA PRESENCIA DE POLIMORFISMOS EN EL METABOLISMO DEL METOTREXATO Y SU EFECTIVIDAD, EN UN GRUPO DE PACIENTES COLOMBIANOS CON ARTRITIS REUMATOIDE", código 2015-025.

5.1. Recolección de información clínica

La población de la que se recolectaría información se estratificó de la siguiente manera:

Población objetivo: Pacientes colombianos con diagnóstico de AR, tratados con MTX y/o ADA

Población accesible: Pacientes adultos con diagnóstico de artritis reumatoide, tratados con MTX y/o ADA, que hayan recibido consulta médica de reumatología en el Hospital Militar Central con información genética disponible para análisis.

Población elegible: Pacientes adultos con diagnóstico de AR, tratados con MTX y/o ADA, que hayan recibido consulta médica de reumatología en el Hospital Militar Central con información genética disponible para análisis y seguimiento por parte del servicio de reumatología que cumplan con los criterios de inclusión.

Criterios de inclusión

- Pacientes que asistieron a consulta de reumatología con diagnóstico de AR y que recibieron tratamiento con MTX como primera línea.
- Pacientes que asistieron a consulta de reumatología con diagnóstico de AR y que recibieron algún DMARD de síntesis química como primera línea y que, por diferentes razones, han comenzado terapia con ADA.

Criterios de exclusión

No existen criterios de exclusión

Selección de la muestra

Muestreo a conveniencia.

Tamaño de la muestra

Todos los pacientes con información genética y clínica disponible que cumplan con los criterios de inclusión. De acuerdo con la investigación de Figueroa et al. (2012), donde estudió el tamaño de muestra requerido para un desempeño eficiente de algoritmos de aprendizaje supervisado, dependiendo del conjunto de datos y del método de muestreo, se necesitaron entre 80 y 560 muestras anotadas para lograr un error medio promedio y un error cuadrático medio inferior a 0,01 (168). Teniendo en cuenta que el número de pacientes con información clínica disponible que cumplen con criterios de inclusión es mayor a 150, se consideró que la cohorte de estudio tenía un tamaño apropiado para la aplicación de estos algoritmos.

Definición de variables

Para este estudio, los desenlaces de efectividad fueron definidos como el puntaje DAS28 del paciente al punto de corte de la toma de información. De acuerdo con la magnitud del puntaje, los pacientes se clasificaron en respondedores (DAS28 < 3,2) y no respondedores (DAS28 ≥ 3,2). Los desenlaces de seguridad fueron definidos como la presencia o ausencia de eventos adversos serios (que justifiquen el cambio de la terapia) durante el periodo de consumo del medicamento.

Se recogieron variables sociodemográficas y clínicas con el fin de caracterizar a la población, así como variables genéticas, que, en conjunto con las variables anteriormente

mencionadas, serían utilizadas como posibles predictores de los desenlaces de efectividad y seguridad. A continuación, se presenta la operacionalización de las variables incluidas en el plan de recolección de datos:

Variable	**Tipo de variable**	**Definición conceptual**	**Definición operacional**
Edad	Cuantitativa Continua	Años vividos del paciente	Años cumplidos a partir de la fecha de nacimiento
Género	Cualitativa Nominal	Sexo biológico del paciente	Femenino o masculino
Edad al diagnóstico	Cuantitativa Continua	Edad del paciente al momento del diagnóstico con AR	Años cumplidos en el momento del diagnóstico con AR
Edad de inicio de tratamiento	Cuantitativa Continua	Edad del paciente al momento del inicio de la terapia antirreumática	Años cumplidos en el momento del inicio de la terapia antirreumática
Estrato	Cualitativa Ordinal	Estrato socioeconómico del paciente	Número que clasifica el estrato socioeconómico
Vivienda	Cualitativa Nominal	Posesión de vivienda por parte del paciente	Presencia o ausencia de vivienda propia
Estado civil	Cualitativa Nominal	Conjunto de circunstancias personales que definen su situación legal en cuanto a su pareja	Condición de una persona en función de si tiene o no pareja
Estado laboral	Cualitativa nominal	Conjunto de circunstancias laborales de una persona	Condición de una persona en función de si tiene trabajo o no

La variable comorbilidad podría ser subdividida a su vez en distintas variables, dependiendo de la cantidad de comorbilidades que presente el paciente. En general, se operacionalizó la variable de esta manera:

Variable	Tipo de variable	Definición conceptual	Definición operacional
Comorbilidad	Cualitativa Nominal	Antecedente de la comorbilidad en el paciente	Presencia o ausencia de antecedente de comorbilidad en el paciente

Los antecedentes toxicológicos se dividieron dependiendo de la información disponible en la historia clínica y se operacionalizaron de la siguiente manera:

Variable	Tipo de variable	Definición conceptual	Definición operacional
Antecedente toxicológico	Cualitativa Nominal	Antecedente de consumo o exposición alguna sustancia que genere dependencia	Presencia o ausencia de antecedente toxicológico en el paciente

Se tuvieron en cuenta 74 polimorfismos relacionados con variabilidad en desenlaces de eficacia y seguridad de MTX e inhibidores del TNF-α, de los cuales adalimumab es un miembro de esa familia. Estas variables se operacionalizaron de la siguiente manera:

Variable	Tipo de variable	Definición conceptual	Definición operacional

Polimorfismo “X” en gen “Y”	Cualitativa Nominal	Valor de la variante genética en el set diploide del individuo	Categorías de homocigotos o heterocigotos para el polimorfismo evaluado

Los desenlaces clínicos de eficacia y seguridad fueron operacionalizados de la siguiente manera:

Variable	**Tipo de variable**	**Definición conceptual**	**Definición operacional**
DAS28	Cuantitativa Continua	Puntaje DAS28 del paciente al momento de la consulta	Valor numérico del DAS28 del paciente
Eventos adversos serios	Cuantitativa Discreta	Eventos adversos que hayan justificado el cambio de terapia farmacológica	Presencia o ausencia de eventos adversos serios
Clasificación funcional	Cuantitativa Discreta	Escala de funcionalidad del paciente al	Valor numérico de la clase funcional del paciente

Las variables relacionadas con la terapia farmacológica fueron:

Variable	**Tipo de variable**	**Definición conceptual**	**Definición operacional**
Fármaco	Cualitativa Nominal	Medicamento antirreumático que consume el paciente al momento de la consulta	Nombre del medicamento que consume el paciente

Dosis	Cuantitativa Discreta	Dosis del medicamento que consume el paciente	Valor en miligramos de la dosis del medicamento
Actual	Cualitativa Binaria	Describe si el paciente consume actualmente el medicamento	Consumo o no consumo actual del medicamento
Motivo de suspensión	Cualitativa Nominal	Causas de la suspensión del tratamiento farmacológico	Categorías de motivo de suspensión: No efectividad, toxicidad, no disponibilidad, otros

5.2. Extracción de ADN y genotipificación

Extracción de ADN

La extracción de ADN fue llevada a cabo utilizando el kit Mo BIO Laboratories. Inc. UltraClean Blood DNA Isolation Kit. El protocolo seguido fue aquel incluido en el kit, que se resume brevemente a continuación:

- Añadir 300 µl de sangre total a 900 µl de la solución G1.
- Invertir dos veces e incubar 5 minutos a temperatura ambiente.
- Invertir dos veces durante la incubación.
- Centrifugar por 30 segundos a 13000 g. Remover el sobrenadante con la punta de una pipeta y descartar sin perturbar la integridad del pellet blanco.
- Resuspender el pellet con el uso de vórtex.
- Revisar la solución G2. Si está precipitada, calentar a 55°-65° por 5 minutos hasta que disuelva. Añadir 300 µl de la solución G2.
- Añadir 1,5 µl de RNasa A, invertir 5 veces y utilizar el vórtex a baja velocidad durante 5 segundos.
- Añadir 100 µl de la solución G3.y utilizar el vórtex en alta velocidad por 15 segundos.

- Centrifugar 3 minutos a 13000 g.
- Transferir el sobrenadante a un tubo colector limpio.
- Agregar 300 µl de Isopropanol al 100%.
- Invertir 15 veces e incubar a temperatura ambiente por 3 minutos.
- Centrifugar 1 minuto a 13000 g.
- Retirar el sobrenadante sin perturbar el pellet y secar con papel absorbente.
- Añadir 100 µl de la solución G4
- Calentar a 65°C en baño de agua hasta resuspensión total del pellet.

Tipificación de SNPs

La lista de SNPs sujetos de genotipificación en los pacientes se escogió de acuerdo con una revisión dirigida de literatura en Pubmed sobre variantes genéticas relevantes en la eficacia y seguridad de tratamientos para artritis reumatoide. De igual manera, se consultó la base de datos PharmacoGenetics KnowledgeBase (PGKB: https://www.pharmgkb.org/) para refinar la lista de SNPs incluidos. La lista de SNPs escogidos tanto para MTX como para ADA se presenta a continuación:

SNPs considerados para MTX
rs1045642, rs1051266, rs11545078, rs1801131, rs1801133, rs2372536, rs3821353, rs4148396, rs4451422, rs4673993, rs4846051, rs4982133, rs4986790, rs5751876, rs5760410, rs6064463, rs6506569, rs6920220, rs70991108, rs719235, rs7279445, rs7499, rs7563206, rs7624766, rs9344, rs9977268
SNPs considerados para anti-TNF
rs1061622, rs1061631, rs10919563, rs11052877, rs12083537, rs1799724, rs1799964, rs1800795, rs1800896, rs1801157, rs1801274, rs1883112, rs20575, rs2229109, rs3397, rs361525, rs3761847, rs3794271, rs3849942, rs394581, rs396991, rs4329505, rs437943, rs4411591, rs4750316, rs548234, rs6028945, rs6071980, rs6138150, rs6427528, rs6691117, rs6822844, rs7046653, rs7527798, rs7574865, rs774359, rs854547, rs854548, rs854555, rs868856, rs928655, rs9514828, rs983332

La tipificación de SNPs se realizó mediante el sistema MassARRAY® de Agena, que detecta las variaciones por medio de espectrometría de masas. Brevemente, este sistema

se basa en la espectrometría de masas de tiempo de vuelo y ionización/desorción láser asistida por una matriz (MALDI-TOF por sus siglas en inglés) y emplea una reacción en cadena de la polimerasa para amplificar las regiones del genoma en las que se encuentra cada SNP. Luego se realiza una reacción de PCR de extensión, en la que un cebador se alinea justo al lado de la base polimórfica. Posteriormente, un nucleótido "terminador" extiende el fragmento de ADN en una base adicional que es específicamente complementaria a la base polimórfica. La base del terminador, que carece de un grupo 3'-hidroxilo, evita que cualquier nucleótido adicional extienda más el fragmento de ADN. Las bases del terminador también tienen su masa modificada, con el fin de que las diferencias de masa entre los fragmentos que difieren en una sola base sean detectables por espectrometría de masas. Así, se puede calcular la masa esperada para el fragmento, que depende de la base polimórfica presente.

Luego de la extensión se agrega un material matricial a cada mezcla de reacción antes de introducirla al espectrómetro. Ya dentro, la muestra es expuesta a pulsos de luz ultravioleta, lo que causa desorción y ionización en la muestra. Finalmente, un campo electrostático de alto voltaje dentro del espectrómetro acelera a las moléculas de ADN ionizadas hacia el detector. Los iones más livianos llegan primero, por lo tanto, luego de cada pulso, el detector mide el tiempo de vuelo relativo de cada analito, que es proporcional a su masa, y de esta manera puede descifrar la base presente en el sitio polimórfico (175).

La información obtenida de la tipificación por MassARRAY® de Agena fue analizada utilizando el software TyperAnalyzer, también propiedad de Agena. El principal filtro de control de calidad de los datos fueron los gráficos de agrupación, realizados por el software utilizando el algoritmo de agrupación *k-means*. De esta manera, los puntos que no pudieron ser correctamente agrupados en homocigotos o heterocigotos fueron excluidos del análisis posterior. Las **Figura 8 y**

Figura 9 esquematizan el proceso de análisis

Figura 8. Esquema de genotipificación con MassARRAY®, cada corrida se va almacenando para dar la gráfica de la derecha. Los puntos rojos esquematizan aquellos que no pudieron ser agrupados en el análisis. Tomado de (176).

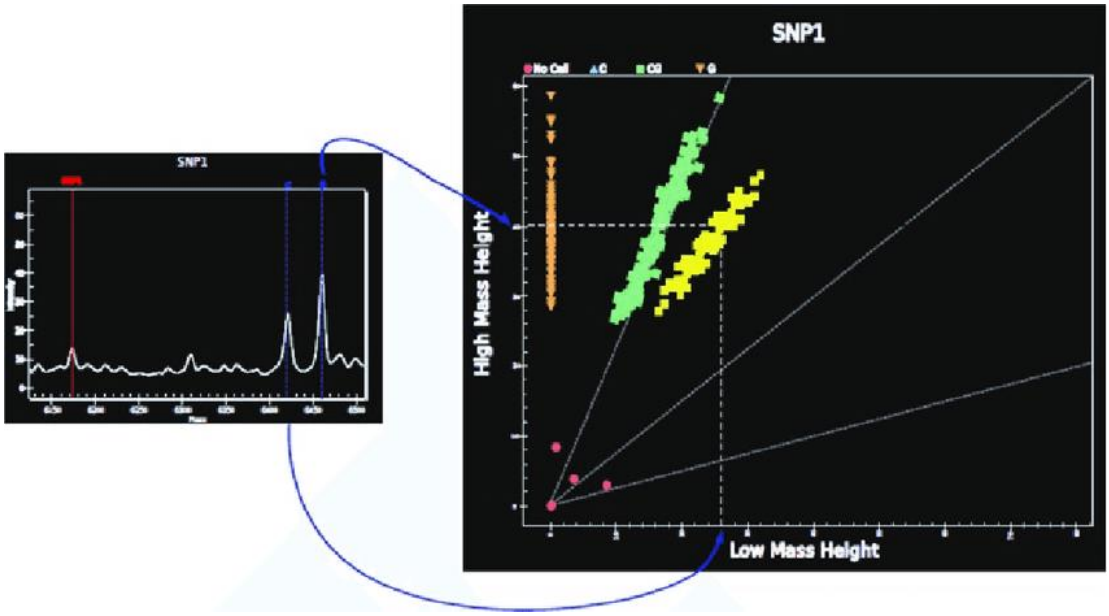
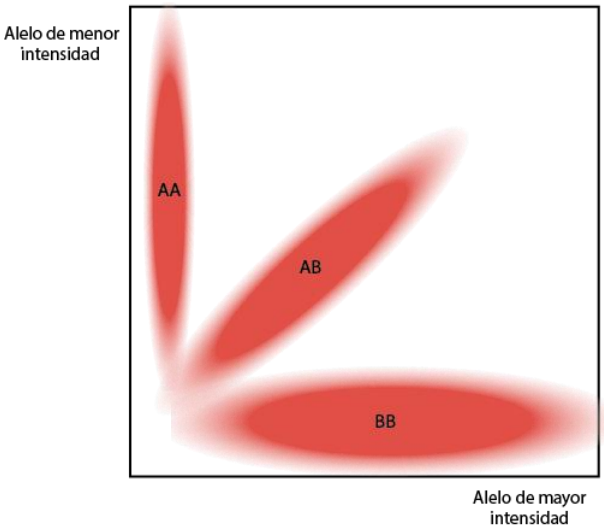


Figura 9. Significado de las nubes de puntos en el llamado de bases. Los tres grupos de genotipos representan las nubes generadas a partir de gráficos de intensidad. El grupo AA consta de todas las llamadas homocigóticas menores, las llamadas heterocigotas corresponden al grupo AB y las llamadas homocigóticas mayores corresponden al grupo BB. Tomado de (177)



5.3. Análisis descriptivo de variables genéticas y clínicas

Limpieza del conjunto de datos

Posterior a la recolección de la información se realizó una limpieza general del conjunto de datos, que consistió en la eliminación de las entradas con más del 50% de información faltante, remoción de caracteres adicionales en los campos como comillas, espacios, puntos y comas que no hacían parte de los valores en cada campo, entre otros caracteres. No se realizó imputación de genes en los casos donde hubo ausencia de reporte del polimorfismo, debido a la incertidumbre de estructura de la población. La información definitiva que sería sujeto de análisis fue almacenada en formato csv. Se excluyeron del análisis las variables con sólo una categoría representada en el conjunto de datos o aquellas que fueran transformaciones de otras variables ya medidas (discretización o categorización posterior).

Descripción de las variables clínicas y genéticas

Las variables clínicas y genéticas recolectadas fueron descritas de acuerdo con su tipo (cuantitativa, cualitativa) usando las medidas de tendencia central para las variables cuantitativas y medidas de frecuencia para las variables cualitativas, la descripción de las variables se muestra en el **Anexo 3**. Para las variables genéticas se realizó un análisis preliminar de equilibrio de Hardy-Weinberg, desequilibrio de ligamiento y se compararon las frecuencias alélicas con aquellas reportadas para la población colombiana (de Medellín) del proyecto 1000 genomas (178,179) y se realizó una prueba de asociación estadística de los distintos SNPs con el desenlace de interés utilizando distintos modelos. Para la evaluación del equilibrio de Hardy-Weinberg se utilizó una prueba chi-cuadrado como se describe en (180). El análisis de desequilibrio de ligamiento, así como las pruebas de asociación estadística se realizaron de acuerdo con lo propuesto en (181,182).

5.4. Selección de variables

Debido a que el conjunto de datos obtenido tiene una alta dimensionalidad, se procedió a realizar un proceso de selección de variables para descartar variables no relevantes y mejorar el poder predictivo de los algoritmos de aprendizaje automático (136). Dado que

no existen marcos de referencia o caminos definidos para la selección de variables (137), se realizó la selección de variables con múltiples algoritmos con el fin de encontrar las variables con mayor influencia individual en el desenlace de interés (respuesta definida por el DAS28). Debido a que la naturaleza de los parámetros clínicos y genéticos es significativamente diferente, se decidió realizar el proceso de selección de variables en el subconjunto de predictores clínicos y por separado en el conjunto de predictores genéticos. Así, se utilizaron los siguientes métodos: *Análisis de correspondencias múltiples*, *prueba de chi-cuadrado, máxima relevancia y mínima redundancia*, *bosques aleatorios*, *eliminación recursiva de características* (RFE por sus siglas en inglés: *Recursive Feature Elimination*) y *regresión logística*. El proceso de selección de variables fue poco factible con el conjunto de datos de ADA debido a la poca cantidad de individuos con información genética disponible en tratamiento con ADA, por lo que se decidió extrapolar algunos de los hallazgos del conjunto de MTX al conjunto de ADA, con el fin de poder aplicar algún algoritmo de aprendizaje automático.

Análisis de correspondencias múltiples

Se inició con este método ya que es un método exploratorio y permite ver la estructura subyacente del conjunto de datos y las categorías de las variables posiblemente más relacionadas con el desenlace de interés. Para esto se utilizó el paquete *FactoMineR* del lenguaje automático R 3.6.3, se excluyeron las observaciones con valores ausentes en alguna variable (n = 27) y se exploró de forma gráfica la capacidad discriminatoria de las variables clínicas en el conjunto de datos. La muestra permaneció representativa de la población de estudio inicial para las medidas consideradas. El número máximo de dimensiones del ACM es de 159 (calculado a partir de la diferencia entre el total de categorías de las variables y el número total de variables – 249 menos 90), que se comparará con el número de dimensiones obtenidas en el análisis.

Prueba chi-cuadrado

La prueba de chi-cuadrado permite establecer asociaciones estadísticas univariadas entre la variable desenlace y los predictores incluidos en el conjunto de datos. En este caso se incluyeron tanto las variables clínicas como las variables genéticas en el análisis, debido a que cada prueba se realiza a un conjunto par de variables y no detecta contribuciones o

ruido que pudieran tener otras variables en el desenlace. Se seleccionaron las variables cuyo valor-p de asociación fuera menor a 0,05 por convención.

Máxima relevancia y mínima redundancia

Para este análisis, se utilizó el paquete *mRMRe* del lenguaje estadístico R 3.6.3. La selección por mRMR se realizó por medio del comando “mRMR.ensemble” que permite la realización del proceso múltiples veces y así considerar el componente estocástico que tiene implícito el cálculo de la información mutua. En cada iteración el algoritmo escogía las 10 variables más relacionadas con el desenlace (*feature_count* = 10). Se realizaron 1000 iteraciones del proceso de selección (*solution_count* = 1000) y se escogieron las primeras 10 variables con mayor frecuencia de selección. Este proceso se realizó para las variables clínicas y genéticas.

Bosques aleatorios

Los bosques aleatorios se utilizaron para calcular la importancia de las variables, por medio de la estimación de la disminución en el índice de Gini para cada variable con 500 árboles de decisión. Esto se realizó utilizando el paquete *randomForest* del lenguaje estadístico R 3.6.3.

Eliminación recursiva de características

La RFE se realizó utilizando el paquete *caret* del lenguaje estadístico R 3.6.3. Debido a que la RFE demanda el uso de otros algoritmos de aprendizaje automático para su ejecución, se decidió utilizar aquellos que venían precargados en el paquete, a saber: bosques aleatorios, árboles de decisión con *bagging* y *Naive Bayes*. Se definieron conjuntos de 1,2,3,4,5,6,7,8,9,10,15,20 y 30 variables como número de variables que podrían retenerse. El método utilizado para crear diferentes conjuntos de entrenamiento y prueba fue la división del conjunto en 10 subconjuntos de entrenamiento y prueba para la validación cruzada (*method* = "repeatedcv", *repeats* = 10). El subconjunto de variables se escogió dependiendo de la exactitud en la clasificación lograda con cada número de variables.

Regresión logística

Para esta aproximación se ajustó un modelo de regresión logística con todas las variables del modelo, de tal forma que se seleccionaran aquellas cuyos coeficientes fueran significativos bajo el criterio del valor-p < 0,05. Se realizó este proceso tanto con las variables clínicas como con las variables genéticas, utilizando la función *glm* del paquete estadístico R 3.6.3.

5.5. Aplicación de algoritmos de aprendizaje supervisado

Dado que el conjunto de datos consta de observaciones etiquetadas con las variables respondedor y no respondedor, se estimó conveniente utilizar algoritmos de aprendizaje supervisado para realizar la predicción. Adicionalmente, teniendo en cuenta que el conjunto de observaciones de ADA estuvo compuesto de pocas observaciones, se decidió implementar únicamente un algoritmo de aprendizaje supervisado: la regresión logística, dado que un algoritmo comúnmente utilizado en epidemiología y se consideró adecuado como estándar de comparación para este conjunto específico de datos. La escogencia de los diferentes algoritmos de aprendizaje respondió a las conclusiones de revisión de literatura, donde los algoritmos más utilizados y con mejor poder predictivo consistieron principalmente de los nombrados a continuación.

Regresión logística

Luego de contar con el conjunto de datos final (aquel obtenido luego del proceso de selección de variables) se ajustó un modelo de regresión logística. El conjunto de datos fue dividido de forma aleatoria en dos grupos: entrenamiento y prueba, el primero constó de 124 observaciones y el segundo de 31 observaciones (una partición 80:20) para este y todos los algoritmos subsiguientes por probar. La función *glm* del paquete estadístico R 3.6.3 fue utilizada para realizar la predicción, ningún parámetro o restricción adicional fue tenida en cuenta para el ajuste de este modelo, dado que su aplicación tiene como finalidad ser el punto de comparación de otros algoritmos. Se realizó un proceso de validación cruzada cambiando las observaciones contenidas en el conjunto de entrenamiento y prueba, con el fin de evaluar la robustez del modelo. El criterio de evaluación fue el área bajo la curva ROC, que permite realizar comparaciones entre clasificadores de forma poco sesgada. Con el fin de mantener la proporción de observaciones en los conjuntos de

entrenamiento y prueba, se realizó un ejercicio de validación cruzada de 5 pliegues o iteraciones (5-fold).

La construcción de la curva ROC para este algoritmo se realizó utilizando el valor de probabilidad de asignación a cada categoría obtenida luego de ajustar la ecuación de regresión (Respondedor o No Respondedor) y cambiando el umbral de clasificación (valor de probabilidad requerido para clasificar cada observación en una categoría u otra).

Árboles de decisión

Se procedió con los árboles de decisión como primer algoritmo de aprendizaje supervisado para la creación de modelos, debido a su simplicidad y a la forma intuitiva y aplicable como presenta los datos. Se utilizó el paquete *rpart* en R para la construcción de árboles, teniendo en cuenta que es una versión actualizada del paquete original *tree* escrito para el lenguaje S. El ajuste de los hiper-parámetros se realizó utilizando el conjunto de entrenamiento. Los hiper-parámetros fueron ajustados con miras al logro del mejor poder predictivo del modelo, medido por el área bajo la curva de las características operativas del receptor. Así, los hiper-parámetros ajustados para el caso de los árboles de decisión fueron:

minsplit: El número mínimo de observaciones que deben existir en un nodo antes de realizar una división.

minbucket: El número mínimo de observaciones que debe tener un nodo terminal. De acuerdo con la documentación del paquete, se consideró adecuado fijar este valor como *minsplit* / 3.

cp: Llamado parámetro de complejidad. Es el factor por el que el subajuste del modelo disminuye al realizar una división; por tanto, cualquier división que no disminuya la falta de ajuste del modelo por un factor igual o mayor a *cp* no va a ser ejecutada. Este parámetro ayuda a ahorrar tiempo de computación al evitar que todas las divisiones posibles sean probadas.

La optimización de los hiper-parámetros fue realizada por medio de una búsqueda exhaustiva en cuadrícula (*exhaustive grid search*), donde los valores de *minsplit* se variaron entre 5 y 50 y los valores de *cp* entre 0 y 1. La profundidad máxima del árbol fue

dejada por defecto (n = 30). Adicionalmente, se realizó un ejercicio de validación cruzada de 5 pliegues (5-fold) para obtener un estimado ponderado de la exactitud del árbol de decisión, dada su alta sensibilidad a la composición de los grupos de entrenamiento y prueba.

La construcción de la curva ROC para los árboles de decisión se realizó utilizando el valor de probabilidad de asignación a cada categoría (Respondedor o No Respondedor) y cambiando el umbral de clasificación (valor de probabilidad requerido para clasificar cada observación en una categoría u otra).

Boosting

Por otra parte, se realizó un ejercicio de *boosting* para el árbol de decisión, utilizando el algoritmo *AdaBoost*, partiendo de *stumps* como clasificadores débiles por defecto, los hiperparámetros de estos *stumps* fueron: *minsplit* = 0, *minbucket* = 2, *cp* = -1, *maxdepth* = 1, *iter* = 500. Brevemente, el algoritmo *AdaBoost* construye una serie de clasificadores débiles de forma secuencial, de tal manera que los errores de clasificación obtenidos por cada clasificador son tenidos en cuenta para la construcción del siguiente clasificador (500 clasificadores sucesivos en total para este análisis). Para este ejercicio se utilizó el paquete *ada* con un algoritmo de *boosting* discreto y una función de pérdida de tipo exponencial, de acuerdo con los valores por defecto del paquete. También se realizó un ejercicio de validación cruzada 5-fold para evaluar la estabilidad del clasificador. La construcción de la curva ROC siguió la misma lógica que aquella construida para el árbol de decisión.

Bosques aleatorios

Como segundo algoritmo de prueba, se procedió con los bosques aleatorios, que pueden concebirse como una extensión de los árboles de decisión, en cuanto los utilizan como base para crear consensos. Este algoritmo resuelve algunos de los problemas de sobreajuste evidenciados en los árboles de decisión, al realizar el proceso de *bagging* (*bootsrap and aggregating*) y realizar la predicción a partir del consenso de múltiples árboles.

Se utilizó la implementación *randomForest* de R para la ejecución del algoritmo. Inicialmente, debido a la ausencia de observaciones en cuatro variables del conjunto de datos (edad, tiempo con AR, rs4846051 y rs3821353), se tuvo que realizar la imputación

de los valores ausentes utilizando el algoritmo de imputación por proximidad, que consiste en la creación del bosque con imputación inicial de los valores ausentes, posteriormente la actualización de los valores ausentes utilizando la proximidad de la información y finalmente, la iteración para el mejoramiento de los resultados. La proximidad de los datos es representada en una matriz de proximidad, una matriz simétrica que describe la frecuencia en la que dos observaciones *i* y *j* llegan al mismo nodo terminal. La proximidad para las variables sin valores ausentes es ponderada y así se escoge el valor de la variable ausente (183).

Se comparó el poder predictivo del árbol construido con el conjunto de datos que tenía valores imputados y aquel construido con el conjunto de datos cuyos valores ausentes fueron removidos. Así mismo, se optimizaron los siguientes hiper-parámetros del algoritmo:

ntree: Se refiere al número de árboles que van a conformar el bosque aleatorio

mtry: Número de variables muestreadas aleatoriamente para construir los conjuntos de datos por *bootstraping*.

maxnodes: Número máximo de nodos terminales que puede tener cada árbol dentro del bosque.

Así mismo, la construcción de la curva ROC para este algoritmo se realizó utilizando el valor de probabilidad de asignación a cada categoría (Respondedor o No Respondedor) y cambiando el umbral de clasificación (valor de probabilidad requerido para clasificar cada observación en una categoría u otra).

Máquinas de soporte vectorial (SVM)

Para la ejecución de este algoritmo, se utilizó la implementación presente en el paquete *e1071* del lenguaje estadístico R 3.6.3. Se probaron *kernels* lineales, polinomiales y radiales con el fin de encontrar aquel que tuviera el mejor poder discriminatorio para este conjunto de datos. Adicionalmente, los parámetros de cada tipo de *kernel* (grado del polinomio y γ para el *kernel* radial), así como el valor de *C* (costo por violación del hiperplano) se optimizaron por medio de una búsqueda exhaustiva con validación cruzada de la siguiente manera: se realizó una validación cruzada de 10 iteraciones (10-fold) para

cada valor de los parámetros mencionados y se escogió el modelo que tuviese el menor error de clasificación. Los parámetros se variaron de la siguiente manera:

- Valor *C*: 0,001, 0,01, 0,1, 1 – 100, 100
- Valor γ: 0,1 – 0,5, 1 – 5
- Grado del polinomio: 2 – 6

Con el *kernel* escogido se realizó un ejercicio de validación cruzada de 5 iteraciones para estimar el área bajo la curva ROC promedio de este algoritmo.

Redes neuronales artificiales (ANN)

Para la ejecución de este algoritmo se utilizó el paquete neuralnet del lenguaje estadístico R 3.6.3. Los siguientes hiper-parámetros fueron sujeto de validación: Función de activación (entre sigmoide y tangente hiperbólica), algoritmo de cálculo de la red (retro-propagación resiliente – *resilient backpropagation*, gradiente estocástico promedio – *Stochastic Average Gradient* y retro-propagación tradicional – *backpropagation*) y arquitectura de la red (número de capas y número de neuronas por capa). Debido a la disponibilidad de poder computacional se probaron arquitecturas de 1 a 3 capas con 1 a 20 neuronas por capa únicamente. La validación se realizó por medio de una búsqueda exhaustiva acoplada a validación cruzada de 5 iteraciones, con el fin de evitar encontrar un conjunto de hiper-parámetros que se sobreajustara al conjunto de entrenamiento.

El desempeño final del algoritmo se estimó como el área bajo la curva ROC promedio de 5 iteraciones de validación cruzada.

6. Resultados

Se consideraron 26 SNPs como posibles predictores de respuesta para MTX y 43 SNPs como predictores de respuesta para anti-TNFs como ADA. Dado el lugar terapéutico de ADA en la práctica clínica colombiana, muy pocos individuos en tratamiento actual con ADA fueron identificados en la base de datos de historias clínicas del Hospital Militar Central. Por lo tanto, los conjuntos finales para análisis consistieron en 155 pacientes con consumo actual de MTX y 12 pacientes con consumo actual de ADA. En el conjunto de datos de MTX el 65% de los pacientes fueron no respondedores y el 35% fueron respondedores bajo el criterio del DAS 28 ≥ 3,2.

Se realizó una prueba de hipótesis de diferencia de medias entre las frecuencias alélicas de la población analizada y la población colombiana analizada en el proyecto de 1000 genomas (178,179). No se evidenciaron diferencias significativas entre las frecuencias alélicas de la mayoría de las variantes para las dos poblaciones. Las frecuencias de las variantes rs10919563, rs1799724 y rs6071980 fueron significativamente diferentes a las de la población del proyecto 1000 genomas.

Equilibrio de Hardy-Weinberg

El equilibrio de Hardy-Weinberg (EHW) es un principio que establece que la variación genética en una población se mantendrá constante de una generación a otra en ausencia de factores perturbadores, de tal manera que las frecuencias de un locus bialélico cualquiera q y p se distribuyen de tal forma que $q + p = 1$ y las frecuencias genotípicas son expresadas por $(q + p)^2$ (184). Desviaciones del EHW en una variante específica pueden indicar que una de las formas de esa variante puede ser deletérea, que hubo una duplicación segmental o que podría haber una asociación del polimorfismo con cierto fenotipo de interés (180). La evaluación del cumplimiento o no de este principio se realiza por medio de una prueba chi-cuadrado de bondad de ajuste, donde los valores p menores

a 0,05 indican una desviación significativa del EHW. Los resultados de esta prueba se muestran en la **Tabla 5**.

Tabla 5. Resultados de la prueba chi-cuadrado de bondad de ajuste para el equilibrio de Hardy-Weinberg. "-" implica que no había representación de algún genotipo en el conjunto de datos, por lo tanto las frecuencias no podían evaluarse.

SNPs considerados para MTX		**SNPs considerados para ADA**	
SNP	**Valor-p**	**SNP**	**Valor-p**
rs1045642	1	rs1061622	0,2199
rs1051266	1	rs1061631	1
rs11545078	-	rs10919563	0,28
rs1801131	0,6055	rs11052877	0,7882
rs1801133	0,1412	rs12083537	0,2695
rs2372536	0,3197	rs1799724	0,7171
rs3821353	0,0066*	rs1799964	1
rs4148396	0,0399*	rs1800795	0,6734
rs4451422	0,8081	rs1800896	0,5632
rs4673993	0,7419	rs1801157	0,3427
rs4846051	-	rs1801274	0,6094
rs4982133	1	rs1883112	0,7954
rs4986790	-	rs20575	1
rs5751876	0,8007	rs2229109	0,1519
rs5760410	1	rs3397	0,5892
rs6064463	0,8063	rs361525	0,1519
rs6506569	0,1046	rs3761847	0,3675
rs6920220	1	rs3794271	0,8058
rs70991108	0,8065	rs3849942	0,6734
rs719235	0,6981	rs394581	0,007*
rs7279445	0,0849	rs396991	0,3915
rs7499	0,6094	rs4329505	-
rs7563206	0,1435	rs437943	0,6053
rs7624766	0,3273	rs4411591	0,0334*

rs9344	0,0087*	rs4750316	0,0042*
rs9977268	0,7419	rs548234	0,0287*
		rs6028945	1
		rs6071980	0,0202*
		rs6138150	0,0086*
		rs6427528	0,041
		rs6691117	0,318
		rs6822844	-
		rs7046653	0,0754
		rs7527798	0,0063*
		rs7574865	0,4387
		rs774359	0,4375
		rs854547	0,626
		rs854548	1
		rs854555	0,6181
		rs868856	0,0754
		rs928655	0,0365*
		rs9514828	0,0523

Tres SNPs mostraron variaciones significativas al EWH en el conjunto de MTX, mientras que 7 SNPs lo hicieron en el conjunto de ADA. Dado el mínimo impacto en la viabilidad del individuo de estas variaciones, es difícil establecer la causa de la desviación.

Desequilibrio de ligamiento

Este término se refiere a la asociación estadística, dentro de los gametos en una población, entre los alelos en dos loci. Aunque el desequilibrio de ligamiento puede deberse a la vinculación, también puede surgir en loci no vinculados; por ejemplo, por selección o por cruces/apareamientos no aleatorios (180). Este análisis ayuda a eliminar variantes que, al estar estadísticamente correlacionadas, no aportan información nueva a los modelos de aprendizaje automático. La correlación de las variantes se mide por medio del coeficiente de correlación r como se muestra a continuación:

$$r = \frac{D_{AB}}{\sqrt{p_A(1-p_A)p_B(1-p_B)}} \; ; D_{AB} = p_{AB} - p_A p_B$$

donde p_i es la proporción del alelo *i* en la población y p_{ij} es la proporción de individuos con ambos alelos *ij* en la población. Entre más cerca está r del valor de 1, mayor es la correlación de ambas variantes. Las **Figura 10** y **Figura 11** muestran el análisis de desequilibrio de ligamiento para los SNPs considerados para MTX y ADA, respectivamente.

Figura 10. Posiciones relativas y resultados del análisis de desequilibrio de ligamiento para los SNPs considerados para MTX.

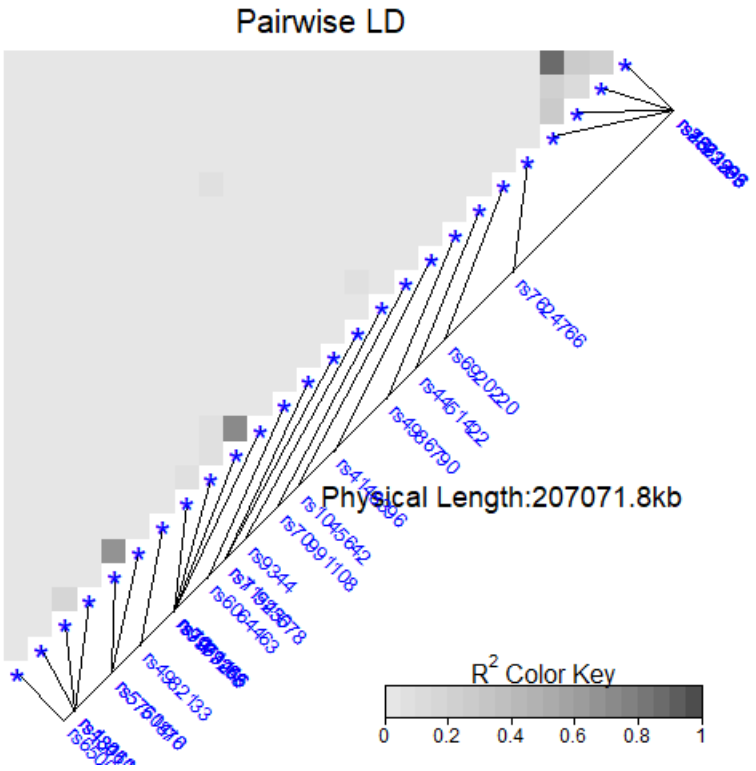
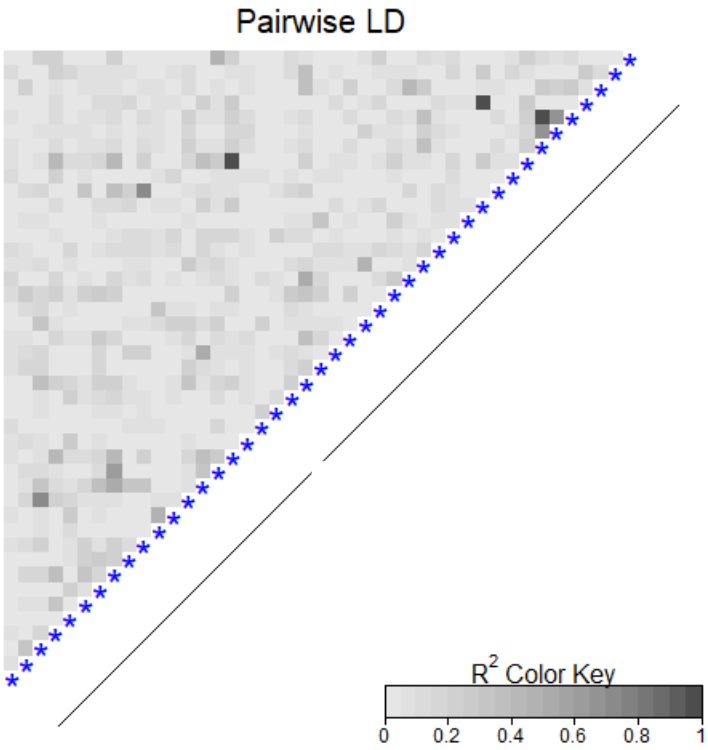


Figura 11. Posiciones relativas y resultados del análisis de desequilibrio de ligamiento para los SNPs considerados para ADA.



Aquellos SNPs con un r^2 mayor a 0,85 fueron considerados como en desequilibrio de ligamiento. Para el conjunto de SNPs relacionados con MTX, aquellos en desequilibrio de ligamiento fueron rs2372536 y rs4673993. Para el conjunto de SNPs relacionados con ADA, aquellos en desequilibrio de ligamiento fueron rs3849942 y rs774359; rs854547 y rs854555; rs7046653 y rs868856.

Asociación univariada de genes

Una aproximación común en los estudios de asociación de genoma completo es la comparación de la prevalencia de un fenotipo particular (enfermedad o condición) en individuos con alelos normales o *wild type* e individuos con alelos variantes. Los SNPs consisten en un alelo mayor (M) y un alelo menor (m), lo que genera 3 genotipos: dos homocigotos (MM y mm) y un heterocigoto (Mm) (185). Dado que las comparaciones se realizan por medio de tablas de contingencia de 2x2 y existen 3 variables, usualmente se agrupan 2 genotipos para realizar comparaciones, obteniendo así distintos modelos

(supuestos) de comparación: Los modelos dominantes comparan MM versus Mm + mm, los modelos recesivos comparan MM + Mm versus mm. Los modelos sobre-dominantes asumen que el heterocigoto tiene el mayor impacto y compara MM + mm versus Mm. Por otro lado, los modelos codominantes que incluyen modelos aditivos y multiplicativos plantean la hipótesis de que MM, Mm y mm están asociados con el riesgo más bajo, el intermedio y el más alto, respectivamente, o viceversa: el riesgo más alto, intermedio y el más alto, respectivamente (186). Para efectos de este análisis, se consideró al alelo menor (m) como aquel alelo de menor frecuencia en la población. Los resultados del análisis asociativo univariado para los conjuntos de SNPs de MTX y ADA se muestran en las **Tabla 6** y **Tabla 7**, respectivamente. Los valores dentro de las tablas corresponden al valor p de cada comparación con el desenlace de interés (respuesta medida en DAS28).

Tabla 6. Resultados del análisis asociativo univariado con los diferentes modelos genéticos para los SNPs considerados para MTX. "-" implica que no había representación de algún genotipo en el conjunto de datos, por lo tanto las frecuencias no podían evaluarse.

Variante	Codominante	Dominante	Recesivo	Sobredominante	Log-aditivo
rs1045642	0,96573	0,8892	0,87275	0,79213	0,99159
rs1051266	0,86911	0,62452	0,74675	0,82289	0,6006
rs11545078	0,51482	-	-	-	-
rs1801131	0,73408	0,45073	0,6675	0,53429	0,43171
rs1801133	0,37583	0,37238	0,49274	0,16271	0,88676
rs2372536	0,15193	0,34032	0,06127	0,90734	0,13969
rs3821353	0,73296	0,57731	0,82418	0,43125	0,80549
rs4148396	0,11199	0,06387	0,81439	0,03862	0,19863
rs4451422	0,73266	0,5241	0,51208	0,94099	0,4314
rs4673993	0,1419	0,12675	0,10194	0,4958	0,05922
rs4846051	0,70819	-	-	-	-
rs4982133	0,22473	0,33076	0,31247	0,13585	0,66208
rs4986790	0,30906	-	-	-	-
rs5751876	0,1748	0,2489	0,07525	0,75689	0,08264
rs5760410	0,88776	0,81875	0,62835	0,84797	0,6713
rs6064463	0,61712	0,97947	0,3562	0,46337	0,62296

rs6506569	0,46151	0,26178	0,91458	0,24535	0,48254
rs6920220	0,95068	-	-	-	-
rs70991108	0,88109	0,68951	0,91352	0,63048	0,85569
rs719235	0,52051	0,30559	0,43734	0,46846	0,25609
rs7279445	0,11896	0,13176	0,33033	0,03907	0,60203
rs7499	0,51417	0,56994	0,26048	0,79213	0,32001
rs7563206	0,42072	0,79639	0,28251	0,25683	0,68192
rs7624766	0,30825	0,42941	0,33076	0,12675	0,92447
rs9344	0,02184	0,041	0,30028	0,00573	0,38584
rs9977268	0,72173	0,47228	0,57787	0,65655	0,42086

Tabla 7. Resultados del análisis asociativo univariado con los diferentes modelos genéticos para los SNPs considerados para ADA. "-"Implica que no había representación de algún genotipo en el conjunto de datos, por lo tanto las frecuencias no podían evaluarse

Variante	Codominante	Dominante	Recesivo	Sobredominante	Log-aditivo
rs1061622	1	-	-	-	-
rs1061631	0,25	-	-	-	-
rs10919563	0,50909	-	-	-	-
rs11052877	1	1	1	0,50909	1
rs12083537	0,70606	-	-	-	-
rs1799724	0,50909	-	-	-	-
rs1799964	1	0,50179	1	0,73314	1
rs1800795	1	0,50179	1	1	1
rs1800896	0,6643	1	0,39765	0,50179	0,62818
rs1801157	0,70606	-	-	-	-
rs1801274	0,54545	0,73314	0,50909	0,16558	0,54545
rs1883112	0,6643	1	0,39765	0,50179	0,62818
rs20575	0,33182	0,20455	1	0,50909	0,33182
rs2229109	0,25	-	-	-	-
rs3397	0,50179	-	-	-	-

rs361525	0,39765	-	-	-	-
rs3761847	0,34545	0,18182	0,50909	0,50909	0,34545
rs3794271	0,70606	-	-	-	-
rs3849942	0,50179	-	-	-	-
rs394581	0,73314	-	-	-	-
rs396991	0,29091	0,70606	0,25	1	0,29091
rs4329505	1	-	-	-	-
rs437943	1	0,70606	1	0,73314	1
rs4411591	0,70606	-	-	-	-
rs4750316	1	-	-	-	-
rs548234	0,70606	-	-	-	-
rs6028945	0,16558	-	-	-	-
rs6071980	0,6643	1	0,39765	0,50179	0,62818
rs6138150	0,09091	0,04545	1	0,18182	0,09091
rs6427528	1	1	1	0,50179	1
rs6691117	0,04545	0,73314	0,04545	0,20455	0,04545
rs6822844	0,25	-	-	-	-
rs7046653	0,91656	1	0,70606	0,73314	0,82638
rs7527798	0,70606	-	-	-	-
rs7574865	0,69928	0,73314	0,39765	0,73314	0,49128
rs774359	0,50179	-	-	-	-
rs854547	0,6643	1	0,39765	0,50179	0,62818
rs854548	1	0,73314	1	0,50179	1
rs854555	0,6643	1	0,39765	0,50179	0,62818
rs868856	0,91656	1	0,70606	0,73314	0,82638
rs928655	0,69928	0,73314	0,39765	0,73314	0,49128
rs9514828	0,31818	0,3115	0,39765	0,20455	0,31818
rs983332	1	0,50909	1	1	1

Utilizando el criterio de valores p bajos en más de un modelo, para el conjunto de SNPs relacionados con la respuesta a MTX, se identificaron 2 SNPs con una asociación significativa con el desenlace: rs4148396 y rs9344. En el caso de los SNPs relacionados

con respuesta a anti-TNFs, se identificó un SNP bajo este criterio: rs6691117. No obstante, dada la baja cantidad de individuos en este último conjunto de datos, esta asociación debe interpretarse con cuidado.

6.1. Selección de variables

Análisis de correspondencias múltiples

El ACM arrojó 128 dimensiones para el conjunto de datos, lo que demuestra que las variables individuales no tienen un gran poder discriminatorio. La **Figura 12** muestra el porcentaje de varianza explicado por cada dimensión en el ACM, mientras que la **Figura 13** muestra la contribución de las variables en la construcción de las dos primeras dimensiones. Las dos primeras dimensiones explican el 3,817% y 3,210% de la varianza, respectivamente. En general, las 10 primeras dimensiones explican sólo el 27,414% de la varianza de todo el conjunto de datos, de lo que se puede concluir que no hubo una separación significativa de los datos con esta técnica.

La **Figura 14** muestra las dos primeras dimensiones del análisis y la posición de las variables en cada una. La concentración de todos los puntos cerca del origen indica que la distribución de las categorías de todas las variables es relativamente homogénea. De igual manera, la **Figura 15** muestra la posición de los individuos en las 2 primeras dimensiones, donde también se observa que, en general, los datos están distribuidos de manera homogénea. También se observa que existen pocos individuos cuya distribución de variables es extrema (los individuos 184, 22, 110, 189 y 218 podrían tener una distribución desigual de variables en comparación con todo el grupo).

Finalmente, la **Figura 16** muestra la distribución de las etiquetas “Respondedor” y “No respondedor” en el conjunto total de individuos en las dos primeras dimensiones. Se observa que estas etiquetas tampoco son un buen discriminador del conjunto de datos y que, en general, no existe una variable individual que pueda separar al conjunto de datos, lo que resalta la necesidad de utilizar de algoritmos más complejos que tengan en cuenta la interacción de las variables y su influencia conjunta en la respuesta terapéutica.

Figura 12. Porcentaje de varianza explicado por cada dimensión en el ACM para las variables clínicas

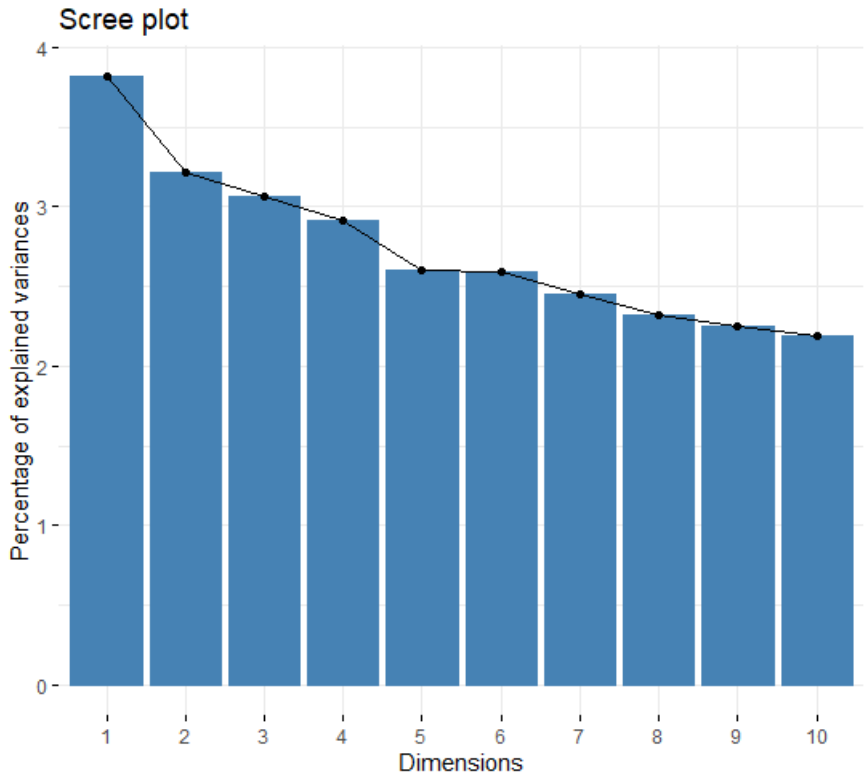


Figura 13. Contribución relativa de las variables en la construcción de las dos primeras dimensiones para las variables clínicas

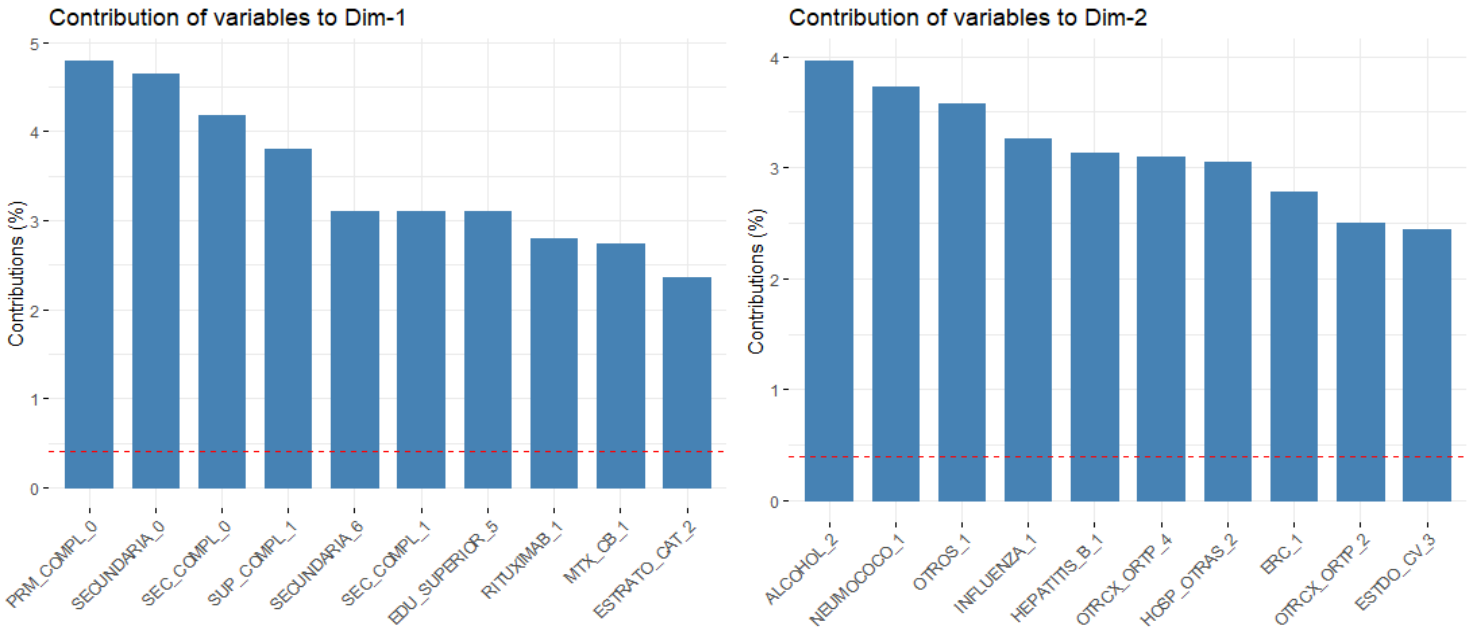


Figura 14. Primeras dos dimensiones del análisis y posición de las variables clínicas en cada una

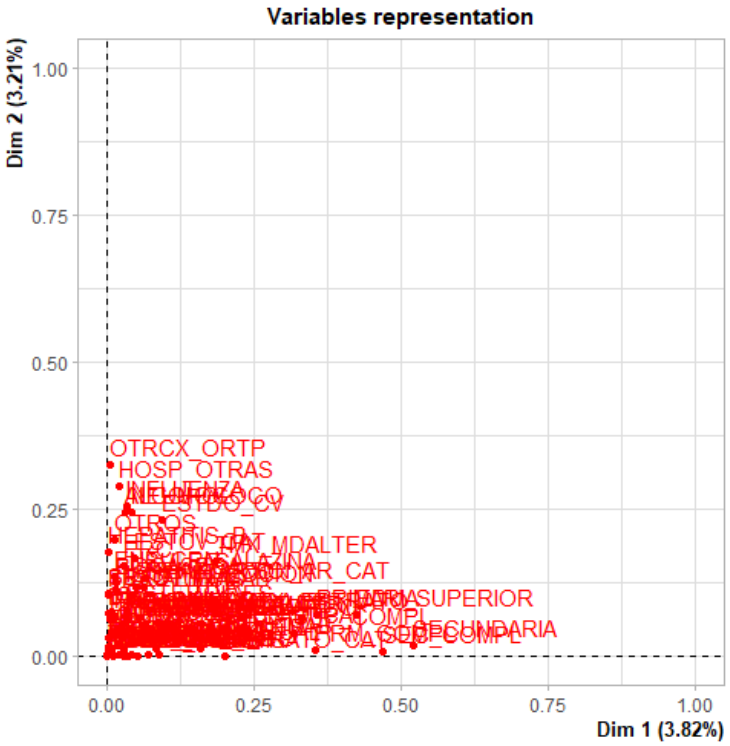


Figura 15. Posición de todos los individuos en las dos primeras dimensiones para el ACM de las variables clínicas

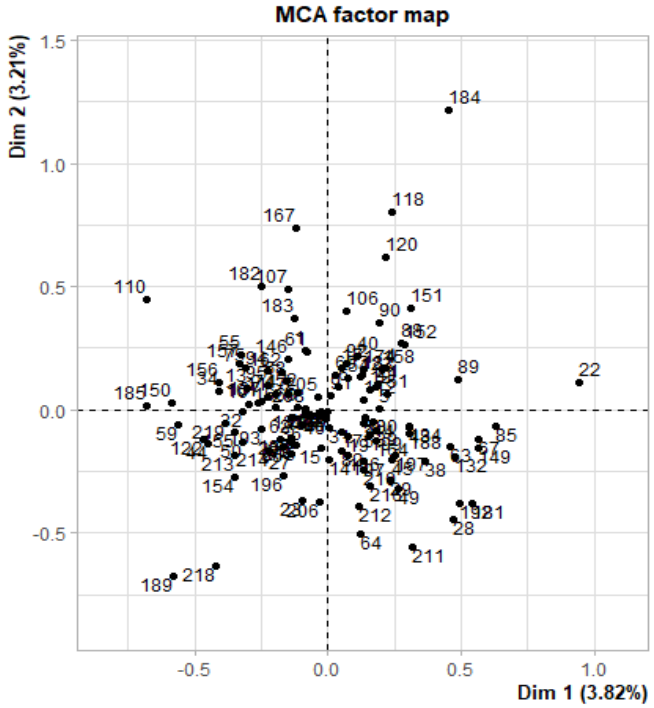
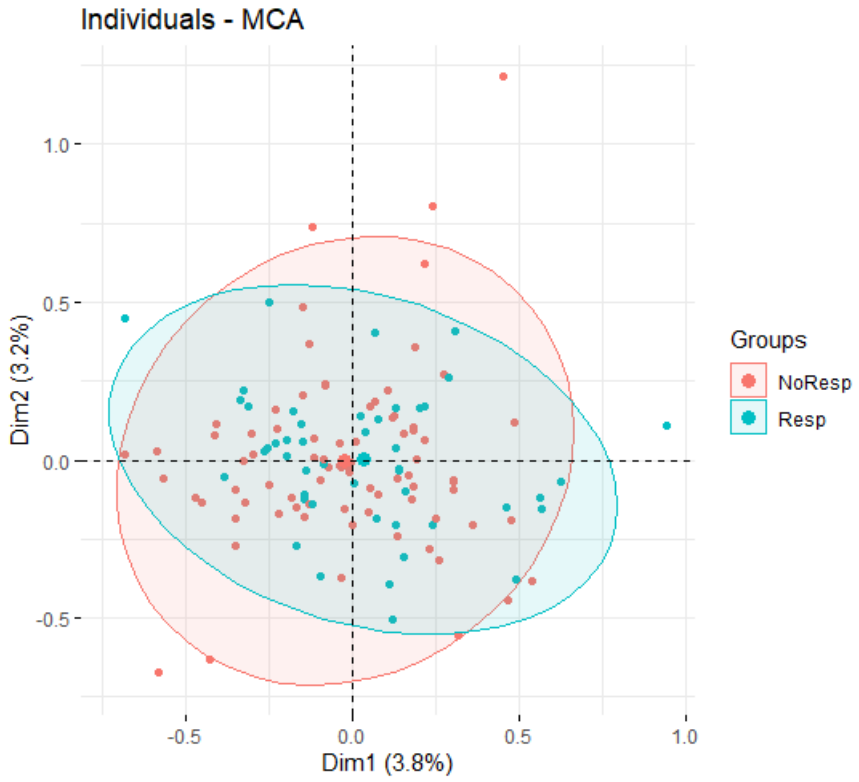


Figura 16. Distribución de las etiquetas "Respondedor" y "No respondedor" de cada individuo en las dos dimensiones generadas con las variables clínicas



Los resultados del ACM para la lista de SNPs del conjunto de datos de MTX se resume en las **Figura 17** a la **Figura 20**. La distribución de genotipos de los SNPs rs4673993 y rs2372536 en los individuos es significativamente diferente a la distribución de los otros SNPs. De igual manera, estos dos SNPs parecen dividir de forma definida a la población; sin embargo, hay que tener en cuenta que, dada la naturaleza exploratoria del análisis, esto no implica que los dos SNPs puedan estar relacionados con el desenlace terapéutico. La capacidad discriminatoria de estas variables podría utilizarse en futuros algoritmos de aprendizaje automático para mejorar su poder de clasificación.

Figura 17. Porcentaje de varianza explicado por cada dimensión en el ACM para las variables genéticas

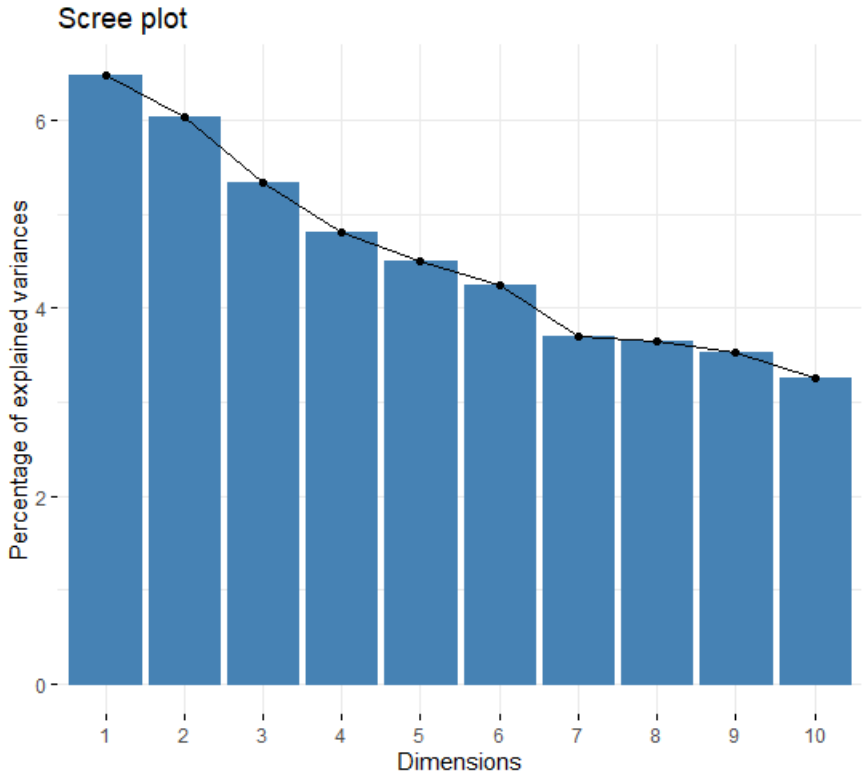


Figura 18. Contribución relativa de las variables en la construcción de las dos primeras dimensiones para las variables genéticas.

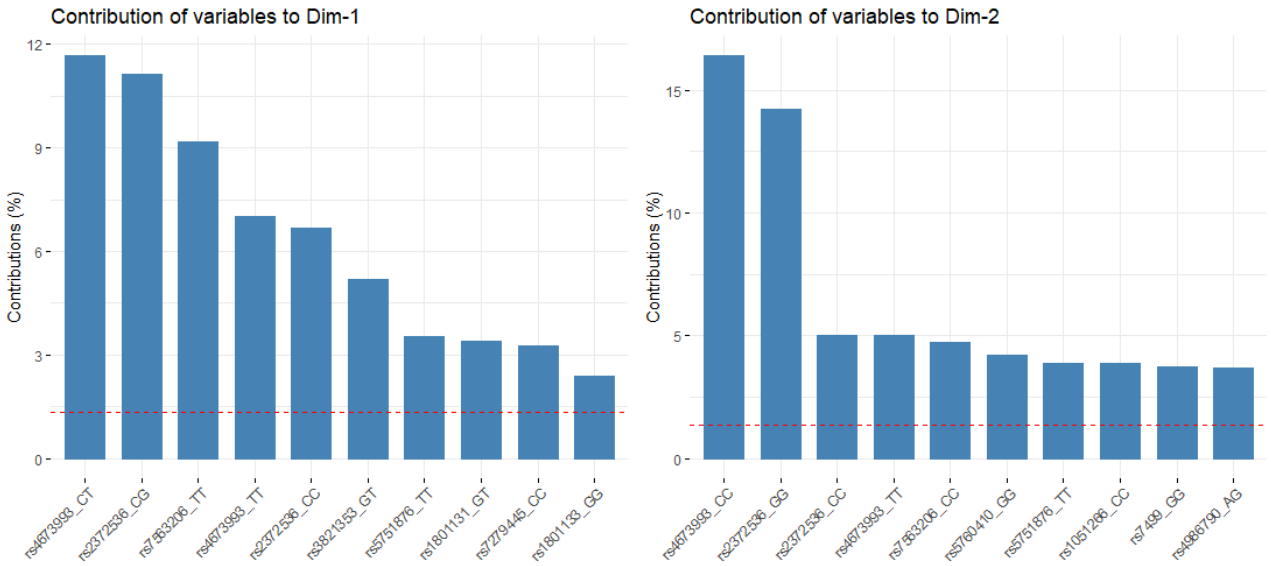


Figura 19. Izquierda: Primeras dos dimensiones del análisis y posición de las variables genéticas en cada una. Derecha: Posición de los individuos en las dos primeras dimensiones para el ACM de las variables genéticas.

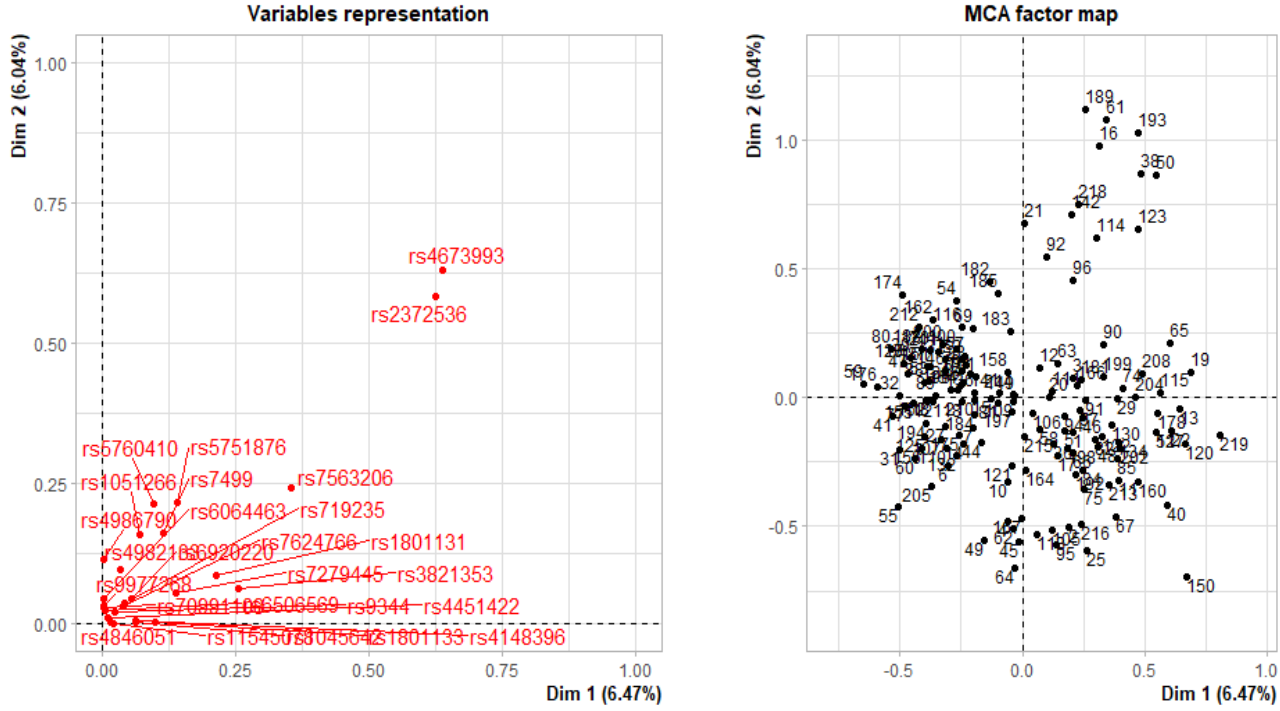
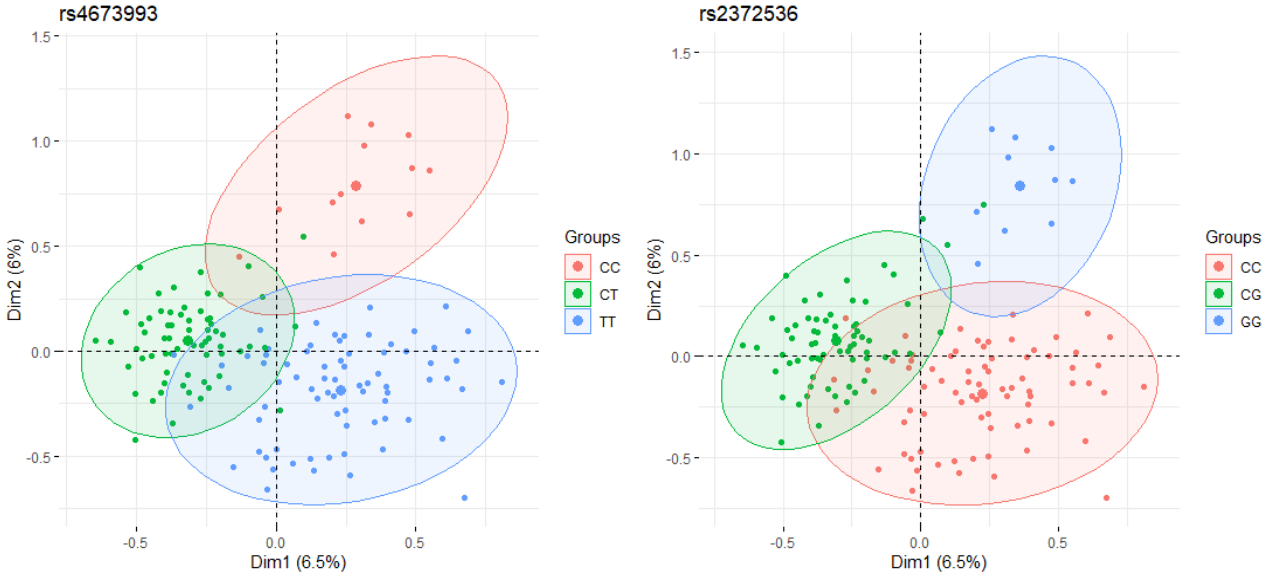


Figura 20. Distribución de los genotipos de los SNPs rs4673993 y rs2372536 de cada individuo en las dos dimensiones generadas con las variables genéticas.



Prueba chi-cuadrado

En la **Tabla 8**, se muestra el resultado de las comparaciones pareadas con valor-p menor a 0,05 en el conjunto de datos de MTX.

Tabla 8. Resultado de la prueba de chi-cuadrado para el conjunto de datos de MTX

Variable	Valor p
Planificación	2,978344e-02
Clasificación funcional	2,937277e-06
rs9344	2,310029e-02

Dos variables clínicas y una variable genética fueron identificadas como significativamente asociadas con el desenlace de la terapia.

Máxima relevancia mínima redundancia

Los resultados del proceso de selección para las variables clínicas se muestran en la **Tabla 9**. Los resultados para las variables genéticas se presentan en la **Tabla 10**. El conteo se refiere a las veces en las que esta variable estuvo dentro de las primeras 10 variables seleccionadas en cada iteración

Tabla 9. Primeras 10 variables clínicas seleccionadas por mRMR

Variable	Conteo
Sequedad genital	166
Planificación	165
Clasificación funcional	163
Vivienda	162
Secundaria	160
Estrato	159
Infliximab	93
Tocilizumab	69
Tiempo con AR	49
Estado civil	35

Tabla 10. Primeras 10 variables genéticas seleccionadas por mRMR

Variable	Conteo
rs4148396	52
rs7279445	52
rs9344	51
rs5751876	50
rs4982133	48
rs7624766	48
rs719235	38
rs1801133	33
rs2372536	32
rs6506569	28

Bosques aleatorios

Las **Figura 21** y **Figura 22** muestran las variables ordenadas descendentemente de acuerdo con su importancia medida como disminución en el índice de Gini en todos los árboles de los que se compone el bosque.

Figura 21. Importancia ordenada de las variables clínicas para el conjunto de MTX

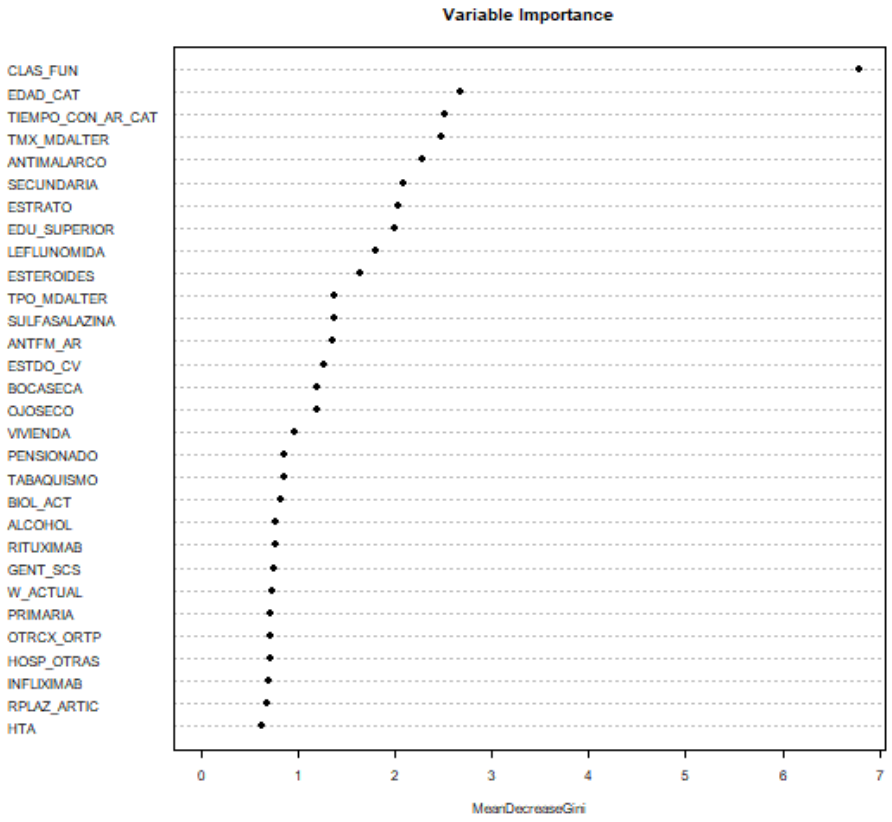
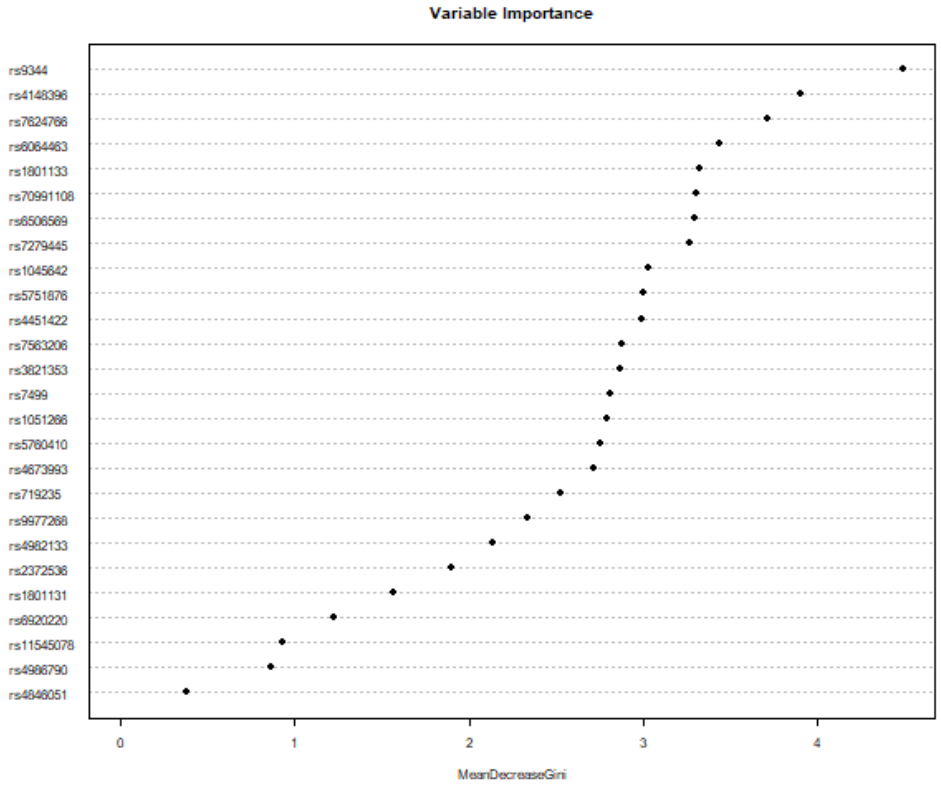


Figura 22. Importancia ordenada de las variables genéticas para el conjunto de MTX



Eliminación recursiva de características

Los resultados del proceso de eliminación recursiva de características para las variables clínicas y genéticas para los 3 métodos de aprendizaje automático se muestran en las **Figura 23** a **Figura 25**. Las variables seleccionadas con cada método se resumen en la **Tabla 11**.

Figura 23. Resultados del proceso de eliminación recursiva de características utilizando los bosques aleatorios como algoritmo base. Izquierda: Variables clínicas. Derecha: Variables genéticas

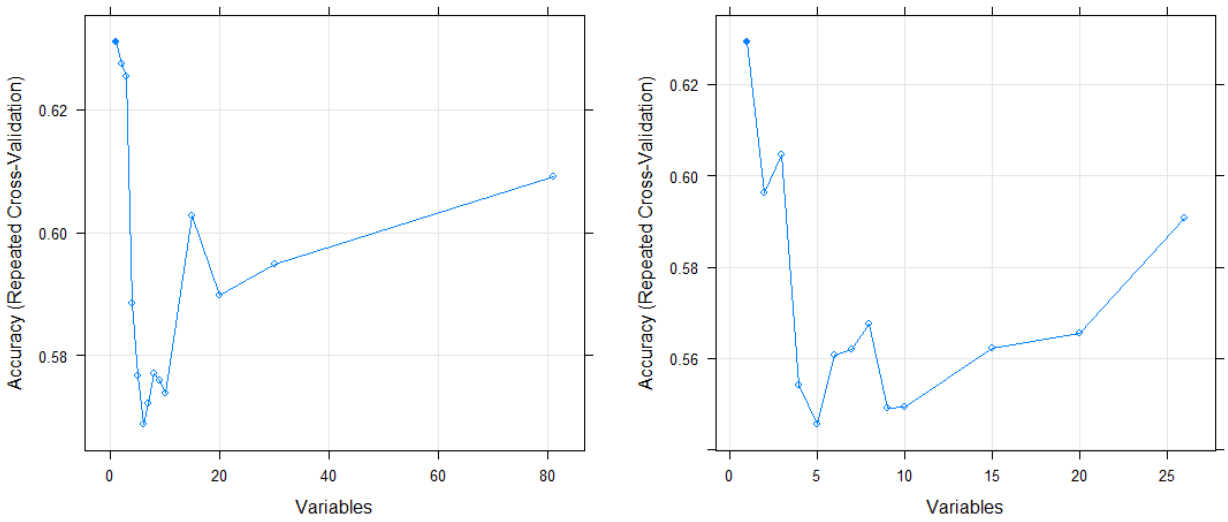


Figura 24. Resultados del proceso de eliminación recursiva de características utilizando *Naive Bayes* como algoritmo base. Izquierda: Variables clínicas. Derecha: Variables genéticas

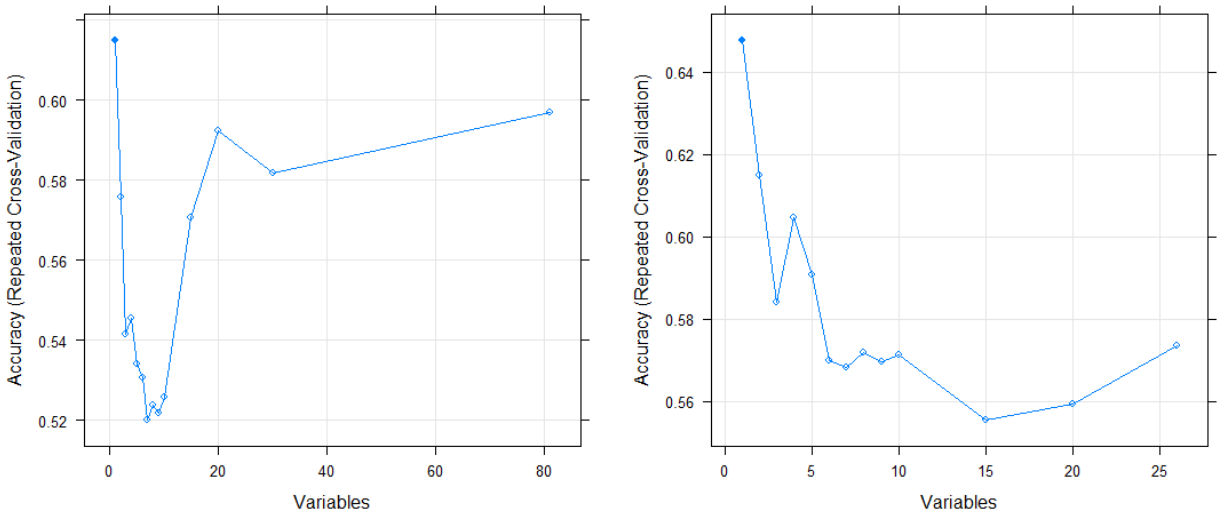


Figura 25. Resultados del proceso de eliminación recursiva de características utilizando árboles de decisión con *bagging* como algoritmo base. Izquierda: Variables clínicas. Derecha: Variables genéticas

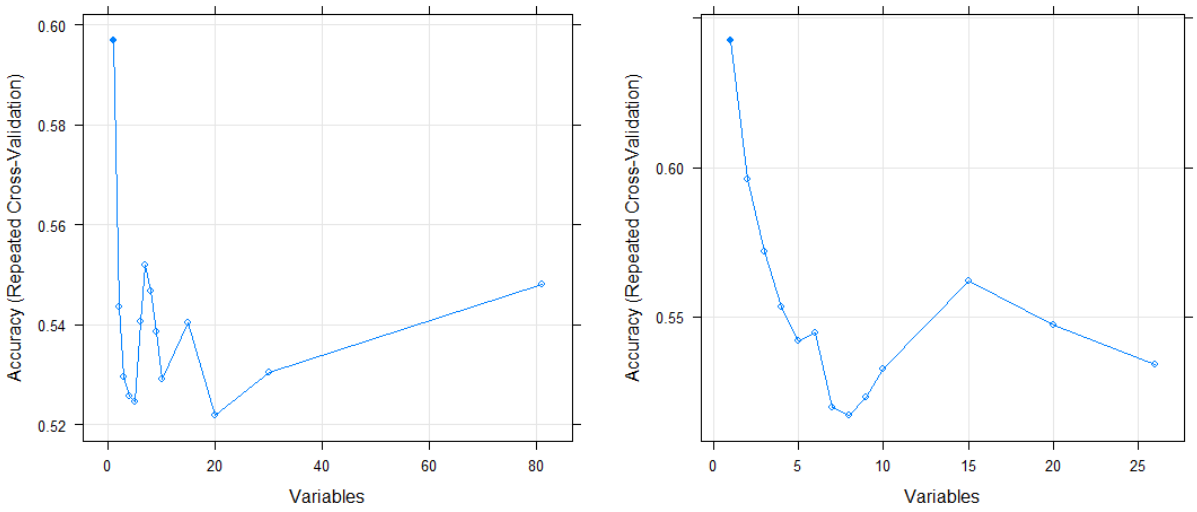


Tabla 11. Variables seleccionadas con cada algoritmo

Algoritmo	Tipo de variable	Variables seleccionadas
Bosques aleatorios	Clínica	Rituximab
	Genética	rs9433
Naive Bayes	Clínica	Antimalárico
	Genética	rs4673993
Árboles de decisión con *bagging*	Clínica	Antecedente de alteración del tratamiento farmacológico
	Genética	rs1801133

Se evidencia que, con todos los métodos de aprendizaje automático, la exactitud de la predicción más alta se logra aparentemente con una sola variable, lo que puede ser un indicador de sobreajuste de los modelos aplicados internamente desde la primera iteración.

Regresión logística

Los coeficientes significativos obtenidos a partir del ajuste de la regresión para las variables clínicas y genéticas se resumen en la **Tabla 12**. El AIC para la regresión con las variables clínicas fue de 218,76 y para la regresión con las variables genéticas 214,98.

Tabla 12. Coeficientes significativos para la regresión logística de variables clínicas y genéticas

Variable	Coeficiente	Valor-p del coeficiente
Estado civil (Otro)	4,133	0,03916
Tiempo con AR (9,2 – 18,4)	-5,169	0,00775
Tiempo con AR (18,4 – 27,6)	-4,160	0,04164
Tabaquismo (antecedente)	3,882	0,03352
Neumococo (Vacunado)	5,276	0,03289
Hospitalización por enfermedad CV (Si)	7,865	0,03291
Consumo concomitante de biológico (Si)	5,348	0,00621
rs4148396 (CT)	1,600	0,01496
rs5751876 (CT)	-2,487	0,01568
rs5751876 (TT)	-6,750	0,00445
rs5760410 (AG)	-2,638	0,01147
rs5760410 (GG)	6,091	0,00776
rs6506569 (CT)	-1,700	0,01168
rs7279445 (CT)	2,547	0,01666
rs7499 (AG)	3,763	0,02940
rs7563206 (CT)	-3,438	0,00737

Listado de variables finales para la ejecución de los algoritmos

En general, se identificaron variables que aparecieron de manera recurrente en los resultados de los algoritmos de selección, por lo que se asume que su impacto en el desenlace de interés es significativo, teniendo en cuenta la plausibilidad biológica de la asociación. Dentro de estas variables se encuentran: rs9433 (RFE, bosques aleatorios,

mRMR y prueba de chi-cuadrado), rs4148396 (bosques aleatorios, mRMR y regresión logística) la clasificación funcional (bosques aleatorios, mRMR y prueba de chi-cuadrado) y el tiempo con AR (regresión logística, bosques aleatorios y mRMR). La inclusión de otras variables clínicas en la lista de predictores se basó tanto en la plausibilidad biológica como en la evidencia de asociación estadística. Para las variables genéticas, debido a que la tipificación de estas en la cohorte se realizó debido a la plausibilidad biológica, se utilizó la asociación estadística como único criterio de inclusión. Se incluyeron en la lista final 9 predictores que se resumen a continuación:

1. Polimorfismo rs9344: Este es un polimorfismo en el gen de la timidilato sintasa (TYMS), una enzima es una enzima metiltransferasa homodimérica que cataliza la reacción de síntesis del timidilato (dTMP). Esta reacción es crucial para la reparación del ADN celular e interactúa en la vía metabólica del MTX (187). Existe evidencia clínica de su asociación con la respuesta a MTX en pacientes con AR (89).
2. Polimorfismo rs4148396: Este es un polimorfismo en el gen del transportador ABCC2 y es una proteína de la familia de los transportadores ABC (*ATP-Binding Cassette*). Las proteínas de la familia ABC transportan diversas moléculas a través de membranas extra e intracelulares (188). Esta proteína en específico es un miembro de la subfamilia MRP que participa en la resistencia a múltiples fármacos y está asociada con la descontinuación de la terapia con MTX en pacientes con AR (189).
3. Polimorfismo rs4673993: Este gen codifica una proteína bifuncional (ATIC) que cataliza los dos últimos pasos de la vía biosintética *de novo* de purinas (190). Existe evidencia clínica que indica que este polimorfismo podría estar relacionado con una mejor respuesta a MTX en pacientes con AR (191,192). Si bien la asociación estadística de este polimorfismo no era tan fuerte, se escogió por su capacidad de separar el conjunto de datos, como se evidenció en el ACM.
4. Polimorfismo rs7279445: Este es un polimorfismo en el gen SLC19A1. La proteína de membrana codificada por este gen es un transportador de ácido fólico y está involucrado en la regulación de las concentraciones intracelulares de ácido fólico (193). El MTX es un antimetabolito del ácido fólico. Polimorfismos en este gen

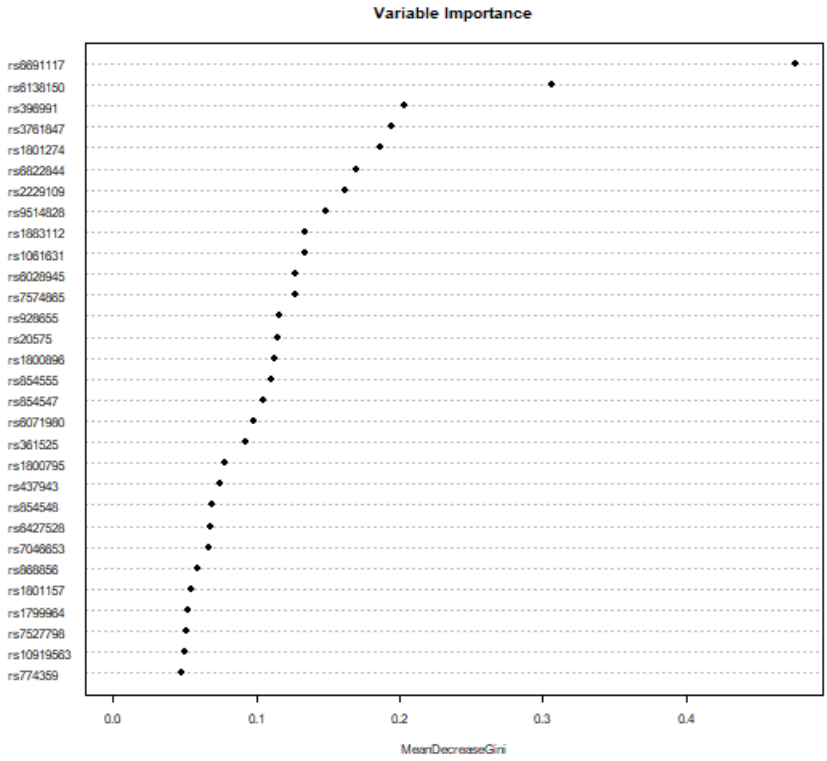
tendrían una asociación con la respuesta terapéutica a MTX en pacientes con AR (194).

5. Polimorfismo rs1801133: Este es un polimorfismo en el gen que codifica a la metilentetrahidrofolato reductasa (MTHFR). Esta proteína cataliza la conversión de 5,10-metilenetetrahidrofolato en 5-metiltetrahidrofolato, un co-sustrato para la remetilación de homocisteína en metionina (195). Esta es una de las dianas directas del MTX. Este polimorfismo podría estar relacionado con la respuesta terapéutica al MTX y a toxicidad (196,197).
6. Clasificación funcional: Se refiere a la clasificación de la capacidad funcional de un paciente con AR, como fue propuesto en (198). Se distinguen 4 clases funcionales:
    I. Capacidad funcional completa para realizar las actividades habituales sin dolor ni limitación.
    II. Capacidad de realizar las actividades habituales a pesar de presentar dolor o limitación en una o más articulaciones.
    III. Capacidad funcional restringida a pocas o ninguna de las actividades o únicamente al cuidado personal.
    IV. Incapacidad. Enfermos confinados en la cama o en una silla.
7. Tiempo con AR: Se refiere al tiempo que llevan los pacientes con el diagnóstico, independientemente del inicio del tratamiento.
8. Edad: Edad del paciente en años
9. Biológicos: Esta variable describe si el paciente consume de forma concomitante medicamentos biológicos

Conjunto de datos de adalimumab

Dado que el conjunto de datos de ADA constó únicamente de 12 observaciones, no se realizó un proceso exhaustivo de selección de variables, debido a que el desempeño de esos algoritmos podría no ser el mejor para ese tamaño de muestra. Se asumió que las variables clínicas como clasificación funcional, edad y tiempo con AR son también predictores de la respuesta a ADA. En cuanto a las variables genéticas, se realizó la selección del top 3 de variables más relacionadas de acuerdo con la reducción en el índice de Gini en los bosques aleatorios como se muestra en la **Figura 26**.

Figura 26. Importancia ordenada de las variables genéticas para el conjunto de datos de ADA



Así, se puede observar que las variables más importantes fueron rs6691117, un polimorfismo en el gen CR1, un gen es miembro de la familia de receptores de activación del complemento (RCA) y está ubicado en la región 'cluster RCA' del cromosoma 1; rs6138150, ubicado en el gen CST5, que hace parte de una superfamilia de inhibidores de proteasas (20) y rs396991, localizado un gen que codifica un receptor para la porción Fc de la inmunoglobulina G, y está involucrado en la eliminación de los complejos antígeno-anticuerpo de la circulación, así como otras respuestas dependientes de otros anticuerpos (199).

6.2. Aplicación de algoritmos de aprendizaje supervisado

Regresión logística

Los parámetros de la regresión obtenida se resumen en la **Tabla 13**.

Tabla 13. Coeficientes obtenidos en la regresión logística final

	Coeficiente estimado	Error estándar	Valor p
Intercepto	-2,24043	3,12279	0,4731
BIOL_ACT1	3,5221	1,07685	0,00107
BIOL_ACT4	12,54583	1455,399	0,99312
CLAS_FUN2	-2,56482	0,80343	0,00141
CLAS_FUN3	-3,65542	1,22388	0,00282
CLAS_FUN4	-2,01522	1,22083	0,0988
rs1801133AG	-1,24663	0,78913	0,11417
rs1801133GG	-0,05726	0,90132	0,94934
rs4148396CT	-1,46377	0,7546	0,0524
rs4148396TT	0,57189	1,18381	0,62903
rs4673993CT	-0,53646	1,2796	0,67504
rs4673993TT	0,47572	1,19148	0,68969
rs7279445CT	2,91107	1,14416	0,01095
rs7279445TT	1,03195	1,13681	0,364
rs9344AG	-1,26154	0,9951	0,20489
rs9344GG	-0,0501	0,98126	0,95928
EDAD_CAT (40,4,52,8]	2,1801	2,59161	0,40023
EDAD_CAT (52,8,65,2]	2,81867	2,46285	0,25243
EDAD_CAT (65,2,77,6]	3,86681	2,49998	0,12193
EDAD_CAT (77,6,90,1]	4,1577	2,5263	0,09981
TIEMPO_CON_AR_CAT (9,2,18,4]	-3,47754	1,1149	0,00181
TIEMPO_CON_AR_CAT (18,4,27,6]	-3,6422	1,20181	0,00244
TIEMPO_CON_AR_CAT (27,6,36,8]	0,69793	1,3807	0,61321
TIEMPO_CON_AR_CAT (36,8,46]	-1,18942	1,78986	0,50635

El criterio de información de Akaike (AIC) de este modelo fue de 126,26, la desviación residual fue 78,261 y la desviación nula 159,526. La matriz de confusión para la predicción con respecto al conjunto de prueba se muestra en la **Tabla 14**.

Tabla 14. Matriz de confusión para la regresión logística

		Predichos	
		No respondedores	**Respondedores**
Reales	**No respondedores**	19	7
	Respondedores	2	3

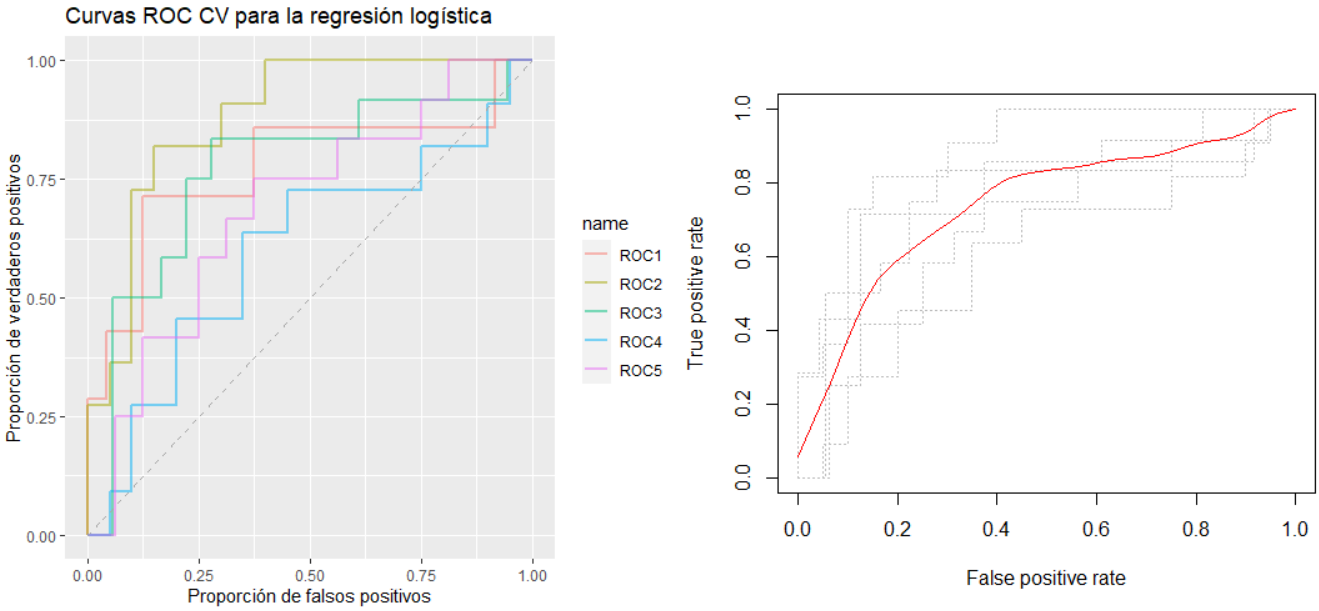
Los resultados de la validación cruzada se muestran en la **Tabla 15**.

Tabla 15. Resultados de la validación cruzada para la regresión logística

Iteración	Exactitud (AUC)
1	0,7738095
2	0,8818182
3	0,7685185
4	0,6000000
5	0,6875000
Promedio simple	**0,7423292**

El área bajo la curva promedio fue de 0,7423292 (95% IC 0,6593611 – 0,8252974), lo que indica un poder predictivo bueno. De igual manera, el desempeño del algoritmo es relativamente consistente durante todas las iteraciones (coeficiente de variación 14,19%). La **Figura 27** muestra las curvas ROC obtenidas para este algoritmo.

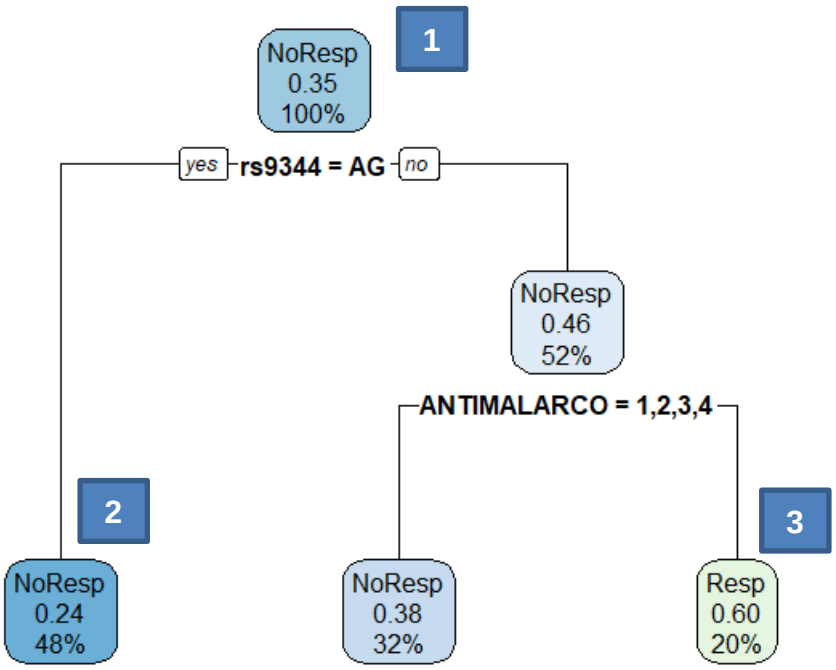
Figura 27. Curvas ROC con validación cruzada para la regresión logística



Árboles de decisión

El valor de los hiperparámetros luego del ajuste fue: *cp* = 0, *minsplit* = 5, *minbucket* = 5

El árbol final obtenido luego del ajuste de los hiperparámetros con el conjunto de datos de entrenamiento se muestra en el **Anexo 4**. En general, la interpretación de los árboles de decisión se muestra a continuación.



1. El 35% de las observaciones en el conjunto de prueba corresponden a pacientes no respondedores; por lo tanto, la probabilidad de ser no respondedor cuando se pertenece a ese 35% es de 1. El nodo inicial contiene al polimorfismo rs9344 y se realiza la división con base en si las bases nitrogenadas en ese sitio polimórfico en el genoma diploide son A y G o no.
2. Si las bases corresponden al polimorfismo heterocigoto AG, se procede al nodo hijo de la izquierda y se observa que el 48% de los pacientes que no respondieron a la terapia con MTX corresponden a pacientes cuya posición polimórfica rs9344 contenía las bases A y G, para una probabilidad de no respuesta de 0,48 con ese factor. El nodo de la derecha (no) corresponde al antecedente de consumo de antimalárico.
3. Si los pacientes con las bases A y G en el sitio polimórfico rs9344 tuvieron antecedente de consumo de antimalárico (valores 1,2,3,4), la probabilidad de responder es de 0,2.

La aplicación de este modelo en el conjunto de prueba dio una exactitud del 74,2%. La tabla de confusión para la predicción se muestra a continuación:

Tabla 16. Matriz de confusión para el árbol de decisión

		Predichos	
		No respondedores	**Respondedores**
Reales	**No respondedores**	16	5
	Respondedores	3	7

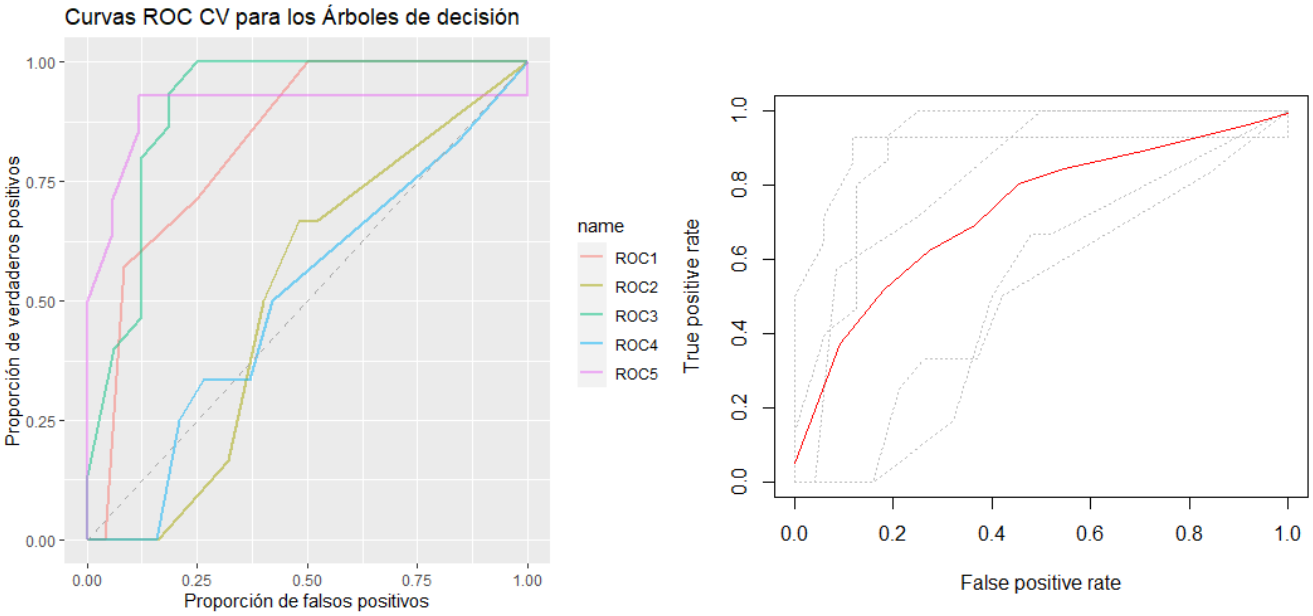
Los resultados de la validación cruzada de 5 iteraciones se resumen en la **Tabla 17**.

Tabla 17. Resultados de la validación cruzada para el árbol de decisión

Iteración	**Exactitud (AUC)**
1	0,8333333
2	0,5133333
3	0,9062500
4	0,5043860
5	0,8991597
Promedio simple	**0,7312925**

Las curvas ROC para las diferentes iteraciones de la validación cruzada se muestra en la **Figura 28**. El área promedio bajo la curva ROC es de 0,731 (95% IC: 0,6491778 – 0,8134071), lo que sugiere un poder predictivo relativamente bueno. Sin embargo, se observa que el modelo es bastante sensible a cambios en el conjunto de datos (coeficiente de variación: 28,04%), lo que estaría indicando sobreajuste en este caso.

Figura 28. Curvas ROC con validación cruzada para los árboles de decisión



Boosting

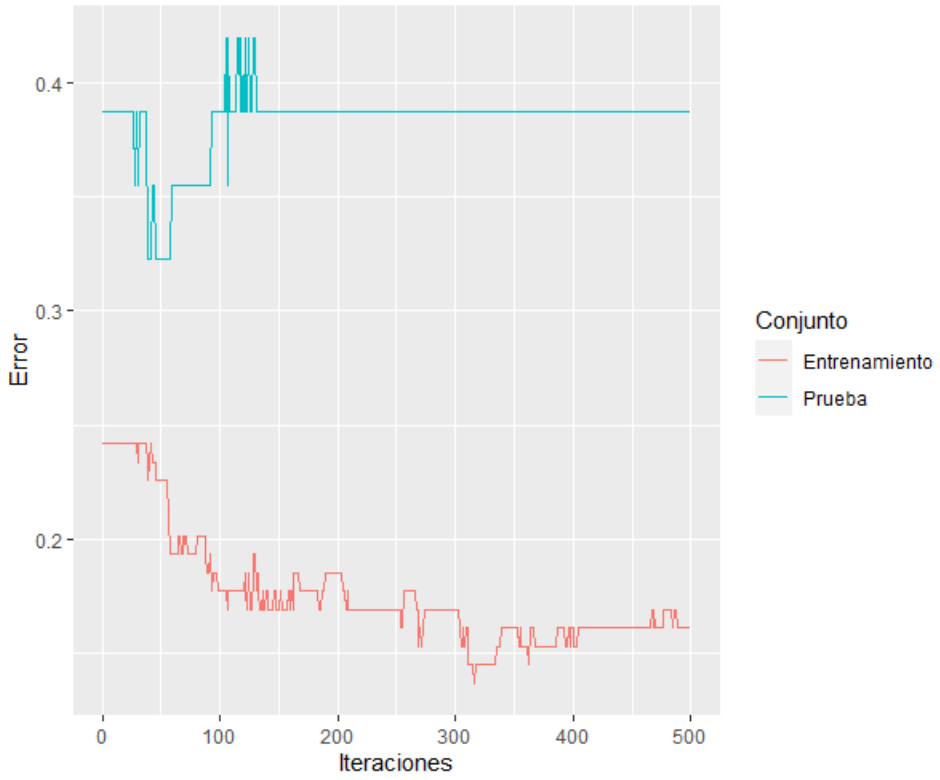
La aplicación de *AdaBoost* con los parámetros ya mencionados supuso un aumento de la exactitud final del clasificador (como área bajo la curva ROC) aumentó ligeramente con respecto al árbol de decisión. Sin embargo, la estabilidad de la clasificación aumentó significativamente, dado que los árboles de decisión sufren de alta varianza, es decir, tienden mucho al sobreajuste. La matriz de confusión para la clasificación del conjunto de prueba con *AdaBoost* se resume a continuación:

Tabla 18. Matriz de confusión para el árbol de decisión con *boosting*

		Predichos	
		No respondedores	**Respondedores**
Reales	**No respondedores**	15	6
	Respondedores	6	4

La tasa de error para los conjuntos de entrenamiento y prueba con cada iteración se muestran en la **Figura 29**. En general, se observa que a medida que aumentan las iteraciones, el error de clasificación disminuye para el conjunto de entrenamiento.

Figura 29. Tasa de error (*misclassification)* como función de las iteraciones de *AdaBoost*

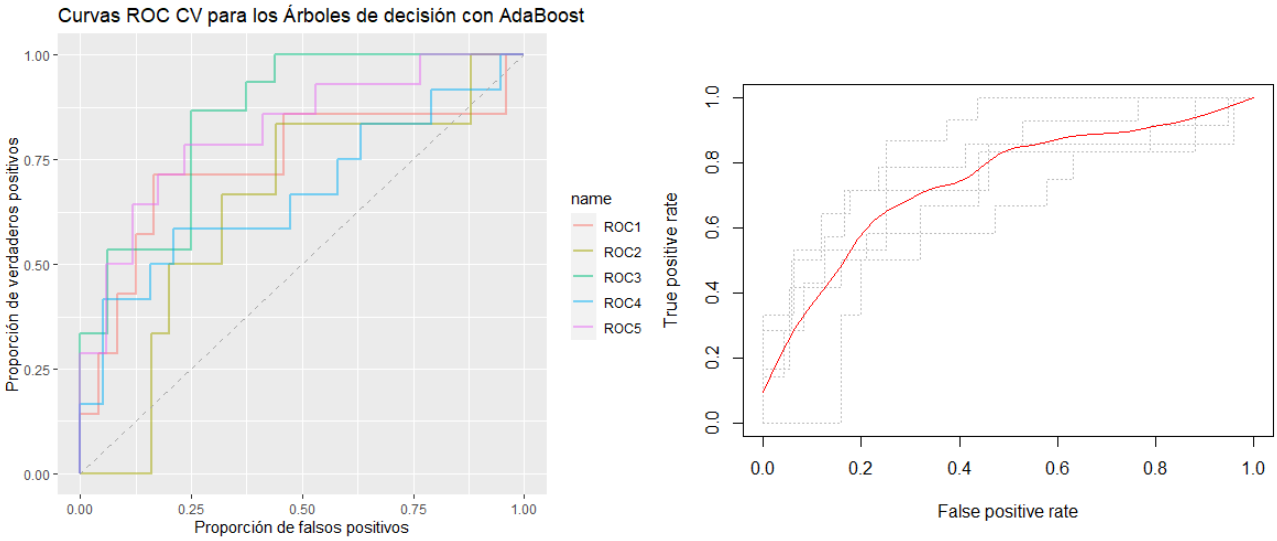


Los resultados de la validación cruzada para *AdaBoost* se muestran en la **Tabla 19**. La **Figura 30** muestra las curvas ROC para cada iteración de la validación cruzada. Se observa que, a diferencia de los árboles de decisión, las curvas ROC para cada iteración son más cercanas entre sí (coeficiente de variación: 12,2%), lo que denota una disminución en el sobreajuste del algoritmo al conjunto específico de entrenamiento con el que fue construido.

Tabla 19. Resultados de la validación cruzada de los árboles de decisión con *AdaBoost*

Iteración	Exactitud (AUC)
1	0,7380952
2	0,6400000
3	0,8500000
4	0,6710526
5	0,8193277
Promedio simple	**0,7436951**

Figura 30. Curvas ROC con validación cruzada para los árboles de decisión con *AdaBoost*



El área bajo la curva ROC promedio para el árbol con *AdaBoost* fue de 0,7436951 (IC 95% 0,6625095 – 0,8248808)

Bosques aleatorios

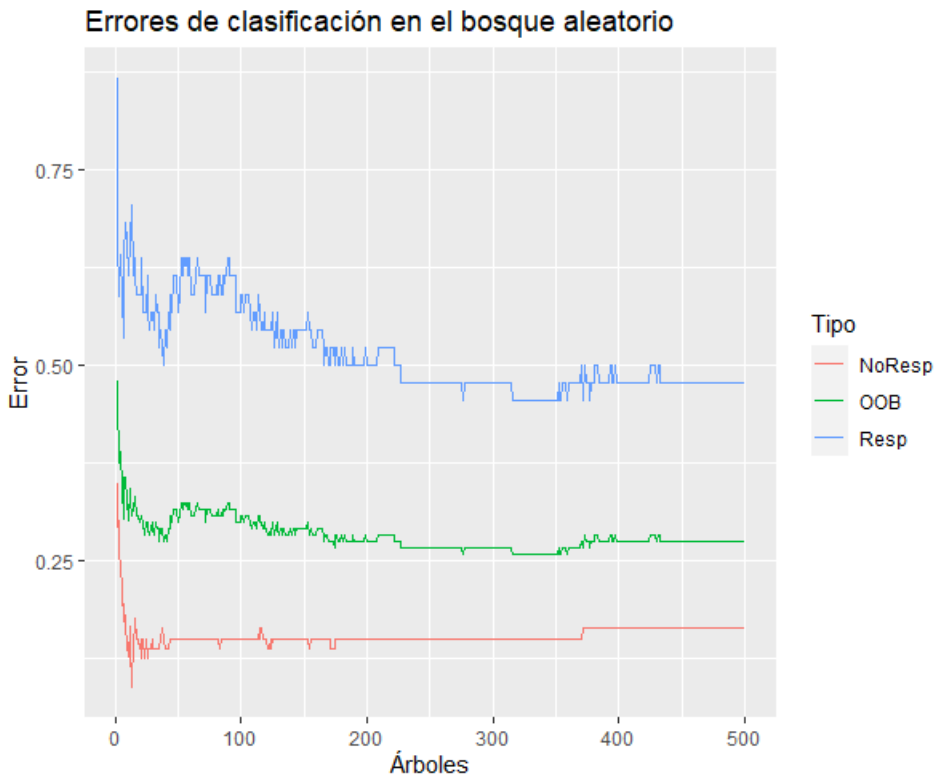
Se realizó la imputación de los valores ausentes en el conjunto de datos con la función *rfImpute*. Se ejecutaron 50 iteraciones hasta que el error OOB (*Out-of-Bag*) se estabilizó en alrededor del 40%. Posteriormente se ajustó el modelo a este conjunto de datos con valores imputados. El proceso de ajuste exhaustivo de los hiperparámetros arrojó los siguientes valores: *ntree* = 500, *mtry* = 1, *maxnodes* = 4.

La tasa de error de clasificación para las dos categorías, así como el error OOB que expresa la proporción de clasificaciones erradas para datos que no fueron incluidos en los conjuntos construidos por *bagging*, se muestran en la **Figura 31**. Se observa que el error OOB convergió en valores alrededor del 25%, mientras que el error de clasificación para los respondedores y no respondedores estuvo cerca del 50% y 12%, respectivamente. La matriz de confusión del bosque obtenido se presenta en la **Tabla 20**.

Tabla 20. Matriz de confusión para los bosques aleatorios

		Predichos	
		No respondedores	**Respondedores**
Reales	**No respondedores**	16	5
	Respondedores	6	4

Figura 31. Errores de clasificación en el bosque como función del número de árboles que lo componen



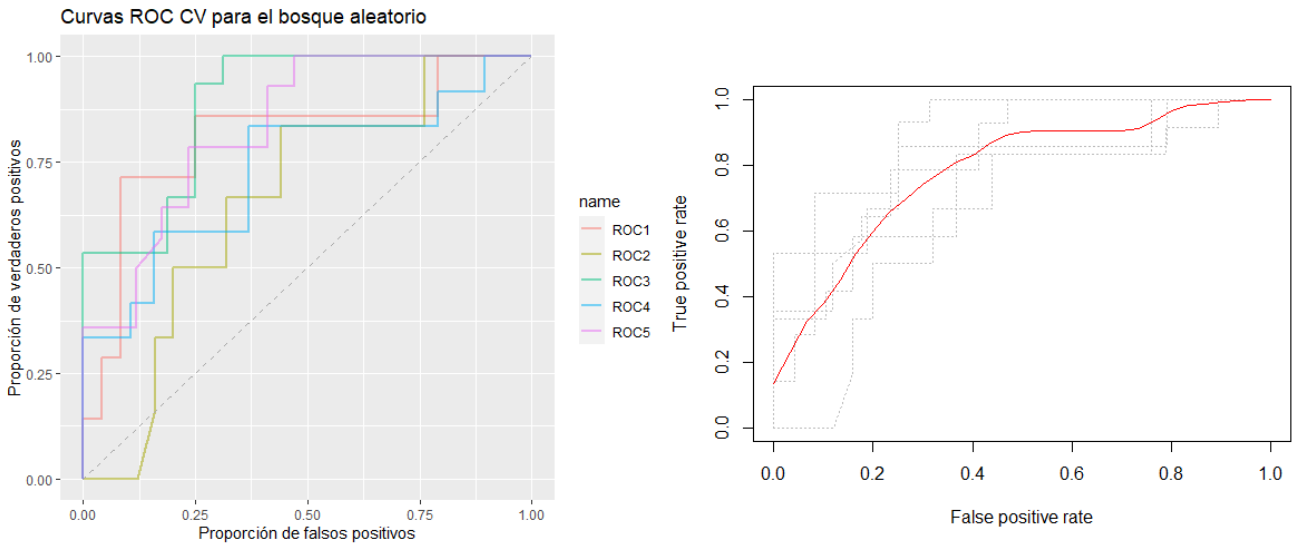
De igual manera, se realizó un proceso de validación cruzada para este algoritmo cuyos resultados se muestran en la **Tabla 21**. Las curvas ROC para cada iteración de la validación cruzada se resumen en la **Figura 32**. Se observa que en general, el poder de clasificación aumentó significativamente con respecto al árbol de decisión y a *AdaBoost*; más aún, la estabilidad de la clasificación se mantuvo en comparación con *AdaBoost* (coeficiente de variación: 11,22%), lo que implica que este algoritmo tampoco sufre de alta

varianza como el árbol de decisión. Finalmente, el área bajo la curva ROC promedio fue de 0,7853694 (95% IC: 0,7119827 – 0,8587560).

Tabla 21. Resultados de la validación cruzada para los bosques aleatorios

Iteración	Exactitud (AUC)
1	0,8095238
2	0,6633333
3	0,8875000
4	0,7324561
5	0,8340336
Promedio simple	**0,7853694**

Figura 32. Curvas ROC con validación cruzada para los bosques aleatorios



Máquinas de soporte vectorial (SVM)

De acuerdo con los ejercicios de optimización de validación cruzada con cada *kernel*, se observó que el *kernel* radial fue aquel con el que se tuvo el menor error de clasificación. Los *kernels* radiales son de la forma:

$$K(x_i, x_{i'}) = \exp\left(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2\right)$$

donde $x_i, x_{i'}$ son puntos correspondientes a dos observaciones diferentes dentro del conjunto de datos. El parámetro γ es una constante positiva cuyo valor óptimo fue encontrado por validación cruzada de 10 iteraciones. Los valores óptimos de γ y *C* fueron 0,1 y 2, respectivamente. La matriz de confusión para el modelo final (aquel con el *kernel* radial y los parámetros γ y *C* escogidos) se muestra en la **Tabla 22**. El desempeño medido como área bajo la curva ROC para la validación cruzada de este modelo se muestra en la **Tabla 23**.

Tabla 22. Matriz de confusión para las SVM

		Predichos	
		No respondedores	**Respondedores**
Reales	**No respondedores**	15	3
	Respondedores	5	7

Tabla 23. Resultados de la validación cruzada para las SVM

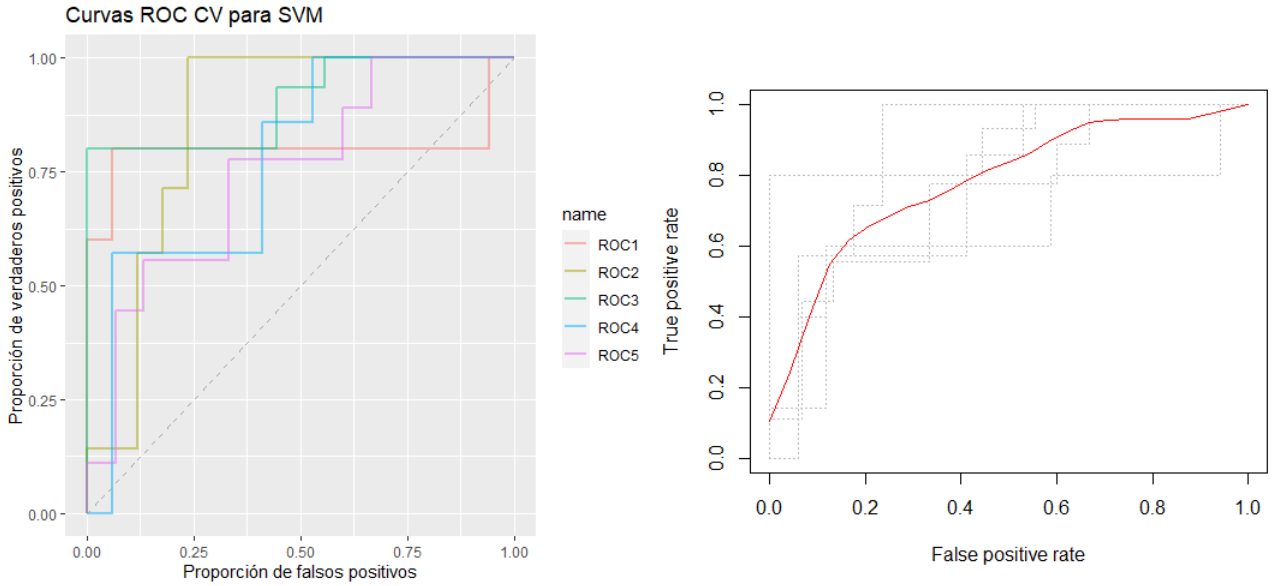
Iteración	**Exactitud (AUC)**
1	0,8000000
2	0,8571429
3	0,9037037
4	0,7731092
5	0,748148
Promedio simple	**0,8164208**

Además, el modelo contó con 93 vectores de soporte, 52 en la categoría de respondedor y 41 en la de no respondedor, lo que indica que un porcentaje alto del conjunto de entrenamiento (75%) contribuyó en la definición del hiperplano discriminante. Si bien esto podría ser un indicador de sobreajuste del modelo, los resultados de la validación cruzada muestran que el valor del área bajo la curva se mantuvo relativamente constante en las

iteraciones (coeficiente de variación: 7,77%), lo que demuestra que el modelo no es excesivamente sensible a los cambios en el conjunto de entrenamiento y, por tanto, se estaría ajustando a la estructura interna general del conjunto de datos.

Finalmente, la **Figura 33** muestra las curvas ROC para las 5 iteraciones. El AUC promedio para las SVM fue de 0,8164208 (95% IC: 0,7376870 – 0,8951546), lo que supone un desempeño mayor que aquel obtenido con los árboles de decisión y el bosque aleatorio.

Figura 33. Curvas ROC para las SVM



Redes Neuronales (ANN)

De acuerdo con los resultados de la búsqueda exhaustiva y validación cruzada dentro de los hiperparámetros seleccionados, los parámetros y arquitectura que resultaron en un mayor poder predictivo medido como el área bajo la curva ROC promedio fueron:

- *Función de activación*: Sigmoide
- *Algoritmo de cálculo de la red*: retro-propagación resiliente con traza de ponderación (*rprop+*)
- *Arquitectura*: 3 capas de 19, 20 y 12 neuronas

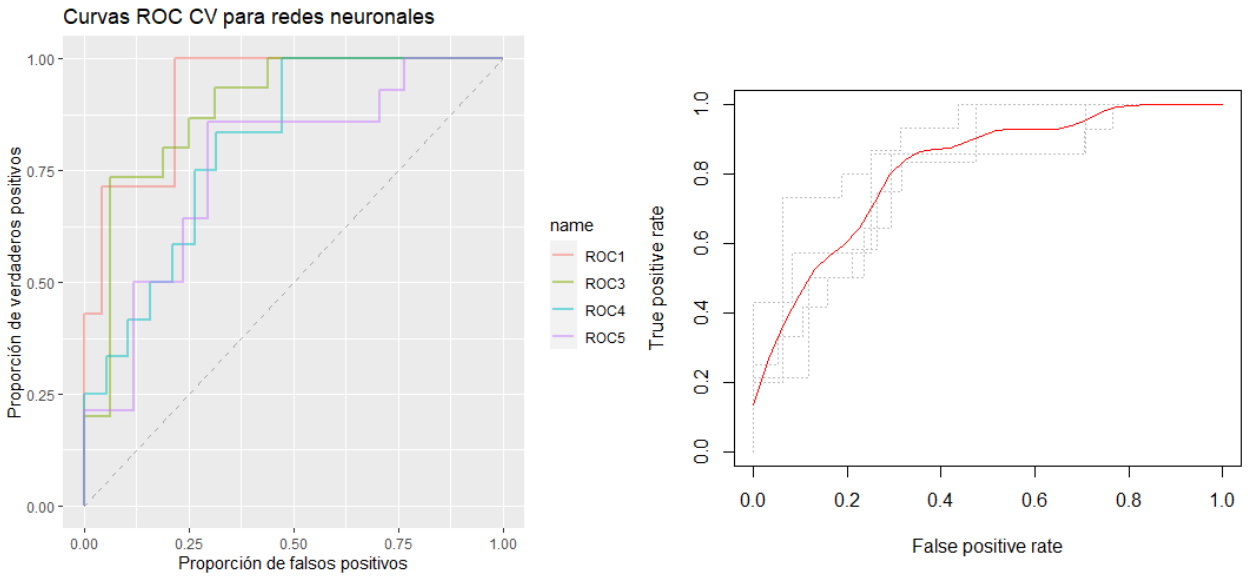
La diferencia entre *rprop+* y la retro-propagación tradicional es el uso de diferentes deltas ponderados para el cálculo del gradiente de la función de costo, lo que lo hace más eficiente en llegar al mínimo de la función que el algoritmo de retro-propagación tradicional.

La **Tabla 24** muestra el desempeño de las redes neuronales a través de todas las iteraciones de validación cruzada y la **Figura 34** muestra las curvas ROC obtenidas para este algoritmo.

Tabla 24. Resultados de la validación cruzada para las redes neuronales

Iteración	Exactitud (AUC)
1	0,9254658
2	0,8175208
3	0,8875000
4	0,8070175
5	0,7647059
Promedio simple	**0,8186749**

Figura 34. Curvas ROC para las redes neuronales



El AUC promedio para las redes neuronales fue de 0,8186749 (95% IC: 0,7460445–0,8913053), lo que supone un desempeño bastante similar al obtenido con las SVM, considerando también que el coeficiente de variación de las ANN fue 7,70%.

Así, a partir de estos resultados se puede evidenciar que los algoritmos con mayor poder predictivo correspondieron a las máquinas de soporte vectorial y las redes neuronales artificiales.

Conjunto de datos de adalimumab

Para el conjunto de datos de adalimumab se realizó una regresión logística, y ningún coeficiente resultó ser significativo en el conjunto de datos, esto pudo deberse principalmente a que el conjunto de datos tenía casi tantas variables como observaciones, lo que dificulta que los algoritmos puedan establecer un modelo que describa la estructura general del conjunto de datos.

7. Discusión

Los algoritmos con mejor desempeño en el conjunto de datos fueron las SVM y las ANN. De igual manera, estos dos algoritmos fueron aquellos con la mayor estabilidad (menor coeficiente de variación) entre las distintas particiones del conjunto de datos. Los algoritmos basados en árboles tuvieron mejor desempeño en general que la regresión logística, que ha sido uno de los algoritmos de clasificación más utilizados en estudios farmacogenéticos en AR (21–23,25). En general, pocos trabajos han comparado el desempeño de los diferentes algoritmos de aprendizaje automático en conjuntos de datos farmacogenéticos y clínicos en artritis reumatoide. Sin embargo, de acuerdo con la revisión realizada, las redes neuronales, las máquinas de soporte vectorial y los bosques aleatorios han sido los que mejor desempeño han tenido en estos conjuntos de datos en distintas enfermedades.

Por otra parte, la mayoría de los algoritmos presentó mayor valor predictivo positivo para los pacientes no respondedores en comparación con los respondedores. Esto podría deberse al desbalance de la distribución de etiquetas en el conjunto de datos (65% no respondedores y 35% respondedores). Sin embargo, dado que en el contexto particular de la enfermedad es mucho más valioso identificar a los potenciales no respondedores que a los respondedores, se consideró a la precisión (o valor predictivo positivo) como el indicador de interés en comparación con la sensibilidad (o *recall*).

Se desarrollaron modelos farmacogenéticos y clínicos para la predicción de desenlaces de efectividad de la terapia con MTX en pacientes con diagnóstico de AR utilizando diferentes algoritmos de aprendizaje supervisado. Aunque se intentó replicar el mismo proceso de descubrimiento de conocimiento con los datos de los pacientes tratados con ADA, dado el tamaño de muestra tan pequeño, resultó imposible extraer conclusiones significativas sobre estos pacientes. No obstante, aunque el número de pacientes con MTX analizados en este trabajo fue significativamente mayor que el de pacientes con ADA, la muestra aún es muy pequeña para muchos de estos algoritmos. El desbalance de información entre los pacientes tratados con MTX y aquellos tratados con ADA es consistente con las proporciones reportadas en otros estudios (46). Los desenlaces de seguridad no pudieron

ser incluidos en el modelo debido a la falta de información detallada en cuanto a su naturaleza y temporalidad.

Aunque la selección de la cohorte se realizó de acuerdo con la disponibilidad de muestra de los pacientes (es decir, que existió un sesgo en la selección), la distribución de la mayoría de alelos para los SNPs considerados no difirieron significativamente de aquella reportada en otros proyectos de tipificación a escala genómica en población colombiana (178,179), lo que es un potencial indicador de representatividad de la población en cuanto a su estructura genética. No obstante, la desviación del EWH de algunas variantes podría indicar un sesgo de selección de la población en estudio (180). De igual manera, se consideraron distintos modelos genéticos para el análisis asociativo univariado entre polimorfismos y desenlace terapéutico; sin embargo, a la hora de aplicar los algoritmos de aprendizaje automático en todo el conjunto de datos, cada valor de la variante genética (homocigoto dominante, heterocigoto y homocigoto recesivo) se tomó como predictor independiente en la construcción del modelo final, lo que evitó la asunción *a priori* de un modelo genético de base. Desafortunadamente, la falta de interpretabilidad de la mayoría de los modelos utilizados impide la evaluación del modelo genético final que fue considerado dentro de cada algoritmo (a excepción de los árboles de decisión y la regresión logística), lo que produce incertidumbre en cuanto a la forma como las variantes genéticas interactúan entre sí en la aparición del desenlace terapéutico.

El proceso de selección de variables previo a la aplicación de algoritmos de aprendizaje automático fue trascendental para este conjunto de datos, debido a que el conjunto de datos crudo poseía más dimensiones que observaciones, lo que hubiera dificultado la generación de algoritmos con buen desempeño en este conjunto. Este problema es bastante común en los proyectos de investigación predictiva de RWD (200,201), lo que demanda una aproximación estandarizada de análisis para este tipo de datos (clínicos y genéticos). De igual manera, dado que esta información fue recogida del contexto clínico real y no como parte de un estudio prospectivo, es común que existan inconsistencias y datos faltantes, por lo que el preprocesamiento es crucial para asegurar datos de buena calidad en los modelos. Desafortunadamente, para este conjunto de datos la información clínica no estaba enteramente disponible en todas las consultas a las que asistió el paciente, por lo que la temporalidad (tiempo con AR) se adicionó como una variable predictora en el modelo para poder dar cuenta de ella en el modelo. Sin embargo,

información intermedia entre los puntos iniciales y finales de medición pudo haber contribuido a una mayor robustez del modelo.

Uno de los supuestos tenidos en cuenta para el desarrollo de este trabajo es que las variantes candidatas encontradas en otras poblaciones tendrían el mismo impacto en la respuesta terapéutica en pacientes colombianos, supuesto que no es verificable con este trabajo, dado que está fuera de su alcance. En este sentido, el trabajo desarrollado por Jenko et al., (2018) (22), en donde el modelo desarrollado en pacientes eslovenos no fue generalizable en pacientes serbios, da cuenta de las limitaciones en la aplicación de estos modelos en grupos de pacientes donde no fueron desarrollados. Estos resultados también indican que pueden existir variables dentro de la estructura poblacional que no se están teniendo en cuenta en los modelos, lo que impide su generalizabilidad. En el caso de este trabajo, si bien la poca información disponible indica que, al menos en cuanto a la distribución de variantes genéticas, la población considerada es similar a la población “de referencia” colombiana, es imposible determinar sin la utilización de otro conjunto de datos si este modelo funcionaría en otra población con una estructura genética similar.

El uso de información de mundo real permite que pueda capturarse la heterogeneidad de los pacientes y así tener más variables en cuenta para la generación del modelo. De igual manera, dado que el objetivo de generación de estos modelos no es encontrar asociaciones novedosas, sino encontrar una manera de agrupar las asociaciones ya encontradas para darle significado clínico, la utilización de evidencia no proveniente de estudios aleatorizados no supone una desventaja para este tipo de estudios.

Por otra parte, dentro del desarrollo de este trabajo se evidenció la significativa necesidad de armonización de historias clínicas en instituciones hospitalarias, debido a que la información utilizada para este estudio provino de fuentes mixtas (texto plano y base de datos), lo que le adicionó complejidad a la estandarización de la información para el análisis final. De igual manera, dado que la recolección de información ómica en el contexto hospitalario no es generalizada en el país, la realización de este tipo de estudios aún se encuentra relegada al ámbito académico. No obstante, como se observó en este trabajo, el potencial uso y aprovechamiento que puede dársele a esta información podría influir significativamente en la toma de decisiones en la práctica clínica colombiana, haciendo que se aprovechen de manera responsable los limitados recursos en salud con los que cuenta el sistema de salud del país.

8. Conclusiones y trabajo futuro

8.1. Conclusiones

A conocimiento del autor, no existen trabajos similares publicados en Colombia sobre el desarrollo de modelos farmacogenéticos en AR, por lo que este trabajo constituye un antecedente para la aplicación de algoritmos de aprendizaje automático y minería de datos en observaciones clínicas y genómicas con miras a su utilización en la práctica clínica. Esto también resalta la necesidad de acoplamiento de proyectos de recolección de información ómica en la práctica clínica colombiana.

8.2. Trabajo futuro

Este trabajo presentó una aproximación analítica a los conjuntos de datos del mundo real con fines predictivos en el contexto colombiano. Dada la heterogeneidad y disponibilidad de datos clínicos localmente, diversas consideraciones deben ser tenidas en cuenta en el preprocesamiento de los datos y en la posterior construcción de modelos. De igual manera, este trabajo permite resaltar los puntos clave que deben ser considerados a la hora de la recolección de datos en el contexto clínico real con el fin de aprovechar de la mejor manera la información farmacogenética y así acelerar su inclusión en la práctica clínica colombiana. Los puntos de trabajo futuro en este campo se encuentran en la inclusión de variables desde diferentes fuentes ómicas (transcriptómica, proteómica, entre otras), así como el desarrollo de modelos explicativos que incluyan el efecto de las variaciones genéticas en la función de las proteínas codificadas (como la velocidad de la reacción, afinidad por el ligando, entre otras) y la relación mecanística de cada uno de los genes incluidos en el modelo.

9. Anexos

A. Anexo 1: Resumen de las estrategias de búsqueda de la revisión sistemática

Reporte de búsqueda # 1

Tipo de búsqueda	Nueva
Bases de datos	PubMed
Plataforma	NCBI
Fecha de búsqueda	11/17/2018
Rango de fecha de búsqueda	Sin restricción
Restricciones de lenguaje	Ninguna
Otros límites	Ninguno
Estrategia de búsqueda (resultados)	("Pharmacogenetics"[Mesh] OR "Genetic Association Studies"[Mesh] OR "Genome-Wide Association Study"[Mesh] OR "Precision Medicine"[Mesh] OR "Pharmacogenomic Variants"[Mesh] OR "Polymorphism, Genetic"[Mesh] OR "Genetic Variation"[Mesh] OR "Models, Genetic"[Mesh]) AND ("Machine Learning"[Mesh] AND "Machine Learning"[All Fields] OR "Statistics as Topic"[Mesh] OR "Medical Informatics"[Mesh] OR "Patient-Specific Modeling"[Mesh] OR "Data Mining"[Mesh] OR "Data Mining"[All Fields] OR "Knowledge Discovery in Databases"[All Fields] OR "KDD"[All Fields] OR "Knowledge Discovery"[All Fields] OR "Logistic Models"[Mesh] OR "Multivariate Analysis"[Mesh] OR "Models, Statistical"[Mesh] OR

	"Discriminant Analysis"[Mesh] OR "Unsupervised Machine Learning"[Mesh] OR "Supervised Machine Learning"[Mesh] OR "Data Mining/classification"[Mesh] OR "Data Mining/methods"[Mesh] OR "Data Mining/statistics and numerical data"[Mesh] OR "Data Mining/trends"[Mesh] OR "Data Mining/utilization"[Mesh]) AND ("Treatment Outcome"[Mesh]) AND ("Drug treatment"[All Fields] OR "Drug Therapy"[All Fields] OR "Medicament therapy"[All Fields] OR "Medicament treatment"[All Fields])
Referencias identificadas	3294

Reporte de búsqueda # 2

Tipo de búsqueda	Nueva
Bases de datos	EMBASE
Plataforma	Elsevier
Fecha de búsqueda	11/17/2018
Rango de fecha de búsqueda	Sin restricción
Restricciones de lenguaje	Ninguna
Otros límites	Ninguno
Estrategia de búsqueda (resultados)	('pharmacogenetics'/exp OR 'genetics, pharmaco' OR 'genetic association study'/exp OR 'genetic association studies' OR 'genomics'/exp OR 'genomics' OR 'personalized medicine'/exp OR 'individualized medicine' OR 'individualized therapy' OR 'individualized medicine' OR 'individualized therapy' OR 'personalized medicine' OR 'personalized therapy' OR 'personalized medicine' OR 'personalized therapy' OR 'precision medicine' OR 'molecular markers'/exp OR 'genetic marker'/exp OR 'genetic markers' OR 'marker, genetic' OR 'dna marker'/exp OR 'dna marker' OR 'marker gene'/exp OR 'gene marker' OR 'gene, marker' OR 'marker gene' OR 'genetic polymorphism'/exp OR 'genetic polymorphism' OR 'polymorphism (genetics)' OR 'polymorphism, genetic' OR 'dna polymorphism'/exp OR 'dna polymorphism' OR 'deoxyribonucleic acid polymorphism' OR 'gene polymorphism' OR 'polymorphism, dna' OR 'single nucleotide polymorphism'/exp OR 'polymorphism, single nucleotide' OR 'single nucleotide polymorphism' OR 'single nucleotide variant' OR 'single nucleotide variation' OR 'genetic variation'/exp OR 'genetic variation' OR 'genome structural variation' OR 'genomic structural variation' OR 'non additive genetic variation' OR 'variation (genetics)' OR 'variation, genetic') AND ('machine learning'/exp OR 'learning machine' OR 'learning machines' OR 'machine learning' OR 'deep learning'/exp OR 'knowledge

	discovery'/exp OR 'statistical learning' OR 'data processing'/exp OR 'automation, computers and data processing' OR 'information processing'/exp OR 'automatic data processing' OR 'computer data processing' OR 'computer processing' OR 'data acquisition' OR 'data collection' OR 'data collection site' OR 'data compression' OR 'data conversion' OR 'data curation' OR 'data display' OR 'data fitting' OR 'data handling' OR 'data logging' OR 'data management' OR 'data processing system' OR 'data recording' OR 'data reporting' OR 'data sampling' OR 'data scanning' OR 'data storage' OR 'data system' OR 'datasets as topic' OR 'date identification' OR 'electronic data processing' OR 'electronical data processing' OR 'focus groups' OR 'idc index system' OR 'information processing' OR 'information processing system' OR 'laboratory data processing' OR 'punched-card systems' OR 'recording, data' OR 'records' OR 'records as topic' OR 'bioinformatics'/exp OR 'bioinformatics' OR 'data mining'/exp OR 'datamining' OR 'big data'/exp) AND ('drug therapy'/exp OR 'drug therapy' OR 'drug treatment' OR 'medicament therapy' OR 'medicament treatment' OR 'medication' OR 'medicinal therapy' OR 'medicinal treatment' OR 'pharmaceutical therapy' OR 'pharmaceutical treatment' OR 'pharmaco-therapy' OR 'pharmaco-treatment' OR 'pharmacological therapy' OR 'pharmacological treatment' OR 'pharmacotherapy' OR 'pharmacotreatment' OR 'therapeutic uses' OR 'therapy, drug' OR 'therapy, pharmacological' OR 'treatment, drug' OR 'treatment, pharmacological') AND ('treatment outcome'/exp OR 'outcome and process assessment (health care)' OR 'outcome management' OR 'patient outcome' OR 'therapeutic outcome' OR 'therapy outcome' OR 'treatment outcome')
Referencias identificadas	2022

Reporte de búsqueda # 3

Tipo de búsqueda	Nueva
Bases de datos	Web of Science
Plataforma	Clarivate Analytics
Fecha de búsqueda	11/17/2018
Rango de fecha de búsqueda	Sin restricción
Restricciones de lenguaje	Ninguna
Otros límites	Ninguno
Estrategia de búsqueda (resultados)	TS=((pharmacogenetics OR "genetic association study" OR "genetic association studies" OR genomics OR "personalized medicine" OR "individualized medicine" OR "individualized

	therapy" OR "individualized medicine" OR "individualized therapy" OR "personalized medicine" OR "personalized therapy" OR "personalized medicine" OR "personalized therapy" OR "precision medicine" OR "molecular markers" OR "genetic marker" OR "genetic markers" OR "dna marker" OR "marker gene" OR "gene marker" OR "genetic polymorphism" OR polymorphism OR "dna polymorphism" OR "deoxyribonucleic acid polymorphism" OR "gene polymorphism" OR "single nucleotide polymorphism" OR "single nucleotide variant" OR "single nucleotide variation" OR "genetic variation" OR "genome structural variation" OR "non additive genetic variation") AND ("machine learning" OR "learning machine" OR "learning machines" OR "deep learning" OR "knowledge discovery" OR "KDD" OR "statistical learning" OR "data processing" OR "information processing" OR "automatic data processing" OR "computer data processing" OR "computer processing" OR "data acquisition" OR "data collection" OR "data collection site" OR "data compression" OR "data conversion" OR "data curation" OR "data display" OR "data fitting" OR "data handling" OR "data logging" OR "data management" OR "data processing system" OR "data recording" OR "data reporting" OR "data sampling" OR "data scanning" OR "data storage" OR "data system" OR "date identification" OR "electronic data processing" OR "electronical data processing" OR "focus groups" OR "information processing" OR "information processing system" OR "laboratory data processing" OR records OR bioinformatics OR "data mining" OR "datamining" OR "big data") AND ("drug therapy" OR "drug treatment" OR "medicament therapy" OR "medicament treatment" OR medication OR "medicinal therapy" OR "medicinal treatment" OR "pharmaceutical therapy" OR "pharmaceutical treatment" OR "pharmaco-therapy" OR "pharmaco-treatment" OR "pharmacological therapy" OR "pharmacological treatment" OR pharmacotherapy OR pharmacotreatment OR "therapeutic uses" OR "drug"))
Referencias identificadas	1799
Sin duplicados	1794

Reporte de búsqueda # 4

Tipo de búsqueda	Nueva
Bases de datos	Scopus
Plataforma	Springer
Fecha de búsqueda	11/17/2018
Rango de fecha de búsqueda	Sin restricción
Restricciones de	Ninguna

lenguaje	
Otros límites	Ninguno
Estrategia de búsqueda (resultados)	(pharmacogenetics OR "genetic association study" OR "genetic association studies" OR genomics OR "personalized medicine" OR "individualised medicine" OR "individualised therapy" OR "individualized medicine" OR "individualized therapy" OR "personalised medicine" OR "personalised therapy" OR "personalized medicine" OR "personalized therapy" OR "precision medicine" OR "molecular markers" OR "genetic marker" OR "genetic markers" OR "dna marker" OR "marker gene" OR "gene marker" OR "genetic polymorphism" OR polymorphism OR "dna polymorphism" OR "deoxyribonucleic acid polymorphism" OR "gene polymorphism" OR "single nucleotide polymorphism" OR "single nucleotide variant" OR "single nucleotide variation" OR "genetic variation" OR "genome structural variation" OR "non additive genetic variation") AND ("machine learning" OR "learning machine" OR "learning machines" OR "deep learning" OR "knowledge discovery" OR "KDD" OR "statistical learning" OR "data processing" OR "information processing" OR "automatic data processing" OR "computer data processing" OR "computer processing" OR "data acquisition" OR "data collection" OR "data collection site" OR "data compression" OR "data conversion" OR "data curation" OR "data display" OR "data fitting" OR "data handling" OR "data logging" OR "data management" OR "data processing system" OR "data recording" OR "data reporting" OR "data sampling" OR "data scanning" OR "data storage" OR "data system" OR "date identification" OR "electronic data processing" OR "electronical data processing" OR "focus groups" OR "information processing" OR "information processing system" OR "laboratory data processing" OR records OR bioinformatics OR "data mining" OR "datamining" OR "big data") AND ("drug therapy" OR "drug treatment" OR "medicament therapy" OR "medicament treatment" OR medication OR "medicinal therapy" OR "medicinal treatment" OR "pharmaceutical therapy" OR "pharmaceutical treatment" OR "pharmaco-therapy" OR "pharmaco-treatment" OR "pharmacological therapy" OR "pharmacological treatment" OR pharmacotherapy OR pharmacotreatment OR "therapeutic uses") AND ("treatment outcome" OR "outcome management" OR "patient outcome" OR "therapeutic outcome" OR "therapy outcome")
Referencias identificadas	3647

B. Anexo 2: Resultados de la revisión sistemática

Autor, año	Población	Información ómica	Predictores	Intervenciones	Desenlaces	Tamaño de muestra	Técnica de clasificación	Hallazgos	Exactitud en la clasificación	Sensibilidad	Especificidad
Rashkin et al., (2019) (103)	Pacientes con cáncer de seno avanzado provenientes de un ensayo clínico aleatorizado que comparaba nab-paclitaxel, paclitaxel e ixabepilona, todos en combinación con bevacizumab. El desenlace medido en el ensayo fué supervivencia libre de progresión	SNPs	Tratamiento, estatus del receptor hormonal, tratamiento previo con taxanos, intervalo libre de enfermedad, presencia de metástasis viscerales y 13 SNPs	Nab-paclitaxel, paclitaxel e ixabepilona, todos en combinación con bevacizumab	Supervivencia libre de progresión	485 pacientes conformaron el conjunto de entrenamiento y 130 el conjunto de prueba	*Elastic Net Regularization*	Se desarrolló un modelo para la predicción de supervivencia libre de progresión utilizando regularización por red elástica, lo que reduce el sobreajuste y realiza la selección de variables. La inclusión de covariables genéticas mejoró la habilidad predictiva del modelo, comparado con uno que sólo usa covariables clínicas.	AUC = 0,81 para el modelo con variables clínicas y genéticas AUC = 0,64 para el modelo sólo con variables clínicas	NR	NR

Autor, año	Población	Información ómica	Predictores	Intervenciones	Desenlaces	Tamaño de muestra	Técnica de clasificación	Hallazgos	Exactitud en la clasificación	Sensibilidad	Especificidad
Naushad et al., (2019) (104)	Pacientes pediátricos con leucemia linfoblástica aguda tratados con el protocolo MCP-841 con tiempo de seguimiento de 2 años	SNPs	Edad, género, 5 polimorfismos en TPMT y 2 en ITPA	6-mercaptopurina	Grado de toxicidad hematológica: 1 (grados 1 y 2), 2 (grados 3 y 4)	96 pacientes: 75% para el conjunto de entrenamiento y 25% para el conjunto de prueba	Árbol de decisión (*CART - Classification and Regression Tree*)	La técnica de aprendizaje automático utilizada pudo discriminar pacientes en cuanto al grado de toxicidad hematológica con 6-mercaptopurina. El modelo tuvo un buen desempeño de acuerdo con el área bajo la ROC	AUC = 0,9649	NR	NR
De Rotte et al., (2018) (24)	Pacientes con artritis reumatoide (AR) pertenecientes a una cohorte prospectiva del estudio tREACH sin tratamiento previo con fármacos antirreumáticos modificadores de la enfermedad (FARME) que comenzaron tratamiento con metotrexate, glucocorticoides con o sin sulfasalazina o hidroxicloroquina.	SNPs	DAS28 de base, HAQ, 2 polimorfismos en ABCB1 y ABCB3, fumador, IMC, folato en eritrocitos	Metotrexato 25mg/semanal	Cambio en el DAS28: >3,2 no respondedor, <3,2 respondedor	270 pacientes para el conjunto de entrenamiento y 84 pacientes para el conjunto de prueba	Regresión logística multivariada	El modelo desarrollado pudo predecir la respuesta insuficiente a metotrexato luego de tres meses de tratamiento, a partir de información clínica y farmacogenética	AUC = 0,80 para las cohortes de entrenamiento y prueba	71%	72%

Autor, año	Población	Información ómica	Predictores	Intervenciones	Desenlaces	Tamaño de muestra	Técnica de clasificación	Hallazgos	Exactitud en la clasificación	Sensibilidad	Especificidad
Maciukiewicz et al., (2018) (106)	Pacientes con desorden depresivo mayor, de ascendencia europea, provenientes de 3 estudios clínicos en los que se administró duloxetina o placebo durante 8 semanas	SNPs	Puntaje *Montgomery–Åsberg Depression Rating Scale* (MADRS) de base, tristeza aparente, reducción del sueño, reducción del apetito, dificultad para concentrarse, inhabilidad para sentir y 19 SNPs	Duloxetina	Respuesta (disminución en el puntaje MADRS mayor al 50%) o remisión (puntaje MADRS igual o menor a 10)	186 pacientes, donde el 20% fue la cohorte de prueba y el 80% la cohorte de entrenamiento	Máquinas de soporte vectorial y árboles de decisión (*Classification Regression Trees* - CRT)	Los modelos desarrollaron tuvieron un desempeño pobre, probablemente debido a la alta tasa de respondedores (70%) en el conjunto de datos inicial. La adición de variables clínicas a los modelos genéticos no aumentó su desempeño	Exactitud = 0,66 para SVM, 0,55 para CRT en el modelo genético para respuesta; 0,51 y 0,45 para SVM y CRT, respectivamente en el modelo genético para remisión	89% y 71% para SVM y CRT, respectivamente en el modelo genético para respuesta 59% y 55% para remisión	9% y 17% para SVM y CRT, respectivamente en el modelo genético para respuesta 41% y 33% para remisión
Lin et al., (2018) (154)	Pacientes con desorden depresivo mayor, de Estados Unidos y Taipei, tratados con inhibidores selectivos de la recaptación de serotonina (SSRI)	SNPs	Puntaje *Hamilton Rating Scale for Depression* (HRSD) al inicio, edad, género, estado marital (o de relación a largo plazo), número de episodios depresivos, intento de suicidio y 10 SNPs	SSRI (escitalopram, paroxetina, duloxetina, citalopram)	Respuesta (disminución en el puntaje HRSD mayor al 50%) o remisión (puntaje HRSD igual o menor a 7)	421 pacientes	*Multilayer Feedforward Neural Networks* (MFNNs)	Los modelos desarrollados superaron en poder predictivo a la regresión logística, sin embargo, se desarrollaron con asociaciones nuevas encontradas en ese estudio, que no tenía suficiente poder estadístico para ello, por lo tanto, su aplicación es limitada.	AUC = 0,82 en promedio para las redes con 1,2 y 3 capas para respuesta a tratamiento y 0,80 en promedio para remisión	75% en promedio para las redes con 1,2 y 3 capas para respuesta a tratamiento y 77% en promedio para remisión	69% en promedio para las redes con 1,2 y 3 capas para respuesta a tratamiento y 66% en promedio para remisión
Zhang et al., (2018) (118)	Pacientes con AR bajo los criterios del American College of Rheumatology (ACR), con duración de síntomas mayor a un año y sin	miRNA, mRNA (basado en chips)	Expresión diferencial de miRNA y mRNA	Tabletas de glicósidos de Trypterisium	Respuesta ACR 20 luego de 12 semanas de tratamiento	43 pacientes: 12 para la cohorte de entrenamiento y 31 para la cohorte de prueba)	Máquinas de soporte vectorial	El modelo logró diferenciar respondedores de no respondedores a partir de cuatro biomarcadores de miRNA. Sin embargo, el modelo estaba enfocado más hacia la validación de los biomarcadores como indicadores indirectos de respuesta, que a evaluar su poder predictivo per se.	AUC = 1, exactitud = 90,32%	NR	NR

Autor, año	Población	Información ómica	Predictores	Intervenciones	Desenlaces	Tamaño de muestra	Técnica de clasificación	Hallazgos	Exactitud en la clasificación	Sensibilidad	Especificidad
	tratamiento previo con FARMEs										
Xie et al., (2018) (162)	Pacientes sometidos a anestesia general para cirugías laparoscópicas, ortopédicas o laparotomía con diferentes agentes anestésicos	SNPs	Características demográficas, tiempo de anestesia, tipo de cirugía, estado físico del paciente de acuerdo con la Sociedad Americana de Anestesiólogos, 20 SNPs	29 fármacos anestésicos y posanestésicos	Tiempo de recuperación, recuperación retardada (>90min)	1453 pacientes	Regresión lineal para el tiempo de recuperación y regresión logística para la recuperación retardada	El modelo farmacogenético para el tiempo de recuperación se ajustó ligeramente mejor a los datos en comparación con el modelo que sólo utilizó variables clínicas. De igual manera, el modelo farmacogenético tuvo un desempeño ligeramente mejor en comparación con el modelo clínico	AUC = 0,76 para el modelo farmacogenético y 0,75 para el modelo clínico en recuperación retardada. AIC = 13229 para el modelo farmacogenético y 13246 para el modelo clínico en tiempo de recuperación	NR	NR
Jenko et al., (2018) (22)	Pacientes adultos con diagnóstico de AR, tratados con metotrexato por al menos seis meses	SNPs	DAS28 de base, erosiones, dosis de metotrexato y 4 SNPs	Metotrexato	Respuesta bajo criterios EULAR (DAS28 < 3,2) luego de seis meses de tratamiento)	110 pacientes eslovenios como cohorte de entrenamiento y 133 pacientes serbios como cohorte de prueba	Regresión penalizada por LASSO	El modelo tuvo un desempeño bueno en la cohorte de entrenamiento, sin embargo, cuando fue utilizado para realizar predicciones en la cohorte de prueba (pacientes serbios), no se desempeñó muy bien	Exactitud = 0,69 en pacientes eslovenos y 0,22 en pacientes serbios	82% en pacientes eslovenos y 8% en pacientes serbios	58% en pacientes eslovenos y 91% en pacientes serbios
Lee et al., (2018) (202)	Pacientes con cáncer gástrico que recibieron cabecitabina + cisplatino con o sin radioterapia (ensayo ARTIST) y fluorouracilo + leucovolina con o	mRNA (basado en chips)	Estado patológico del tumor, estado del nodo, recurrencia, invasión linfovascular, edad, género, expresión diferencial de mRNA	Cabecitabina + cisplatino con o sin radioterapia y fluorouracilo + leucovolina con o sin radioterapia	Sobrevivencia global de 1 a 5 años	1886 pacientes: 80% para el conjunto de entrenamiento y 20% para el conjunto de prueba	*Survival Recurrent Network* (SRN)	La red de sobrevivencia recurrente pudo predecir con exactitud la probabilidad de supervivencia de los pacientes en cada punto del tiempo, siendo la predicción a 5 años la más exacta	AUC = 0,858, 0,869, 0,879, 0,912 y 0,953 para las predicciones para del primer al quinto año, respectivamente	NR	NR

Autor, año	Población	Información ómica	Predictores	Intervenciones	Desenlaces	Tamaño de muestra	Técnica de clasificación	Hallazgos	Exactitud en la clasificación	Sensibilidad	Especificidad
	sin radioterapia (FU/LV/RT)										
Yang et al., (2018) (99)	Pacientes con infección por *Mycobacterium tuberculosis*	SNPs	SNPs relacionados con la resistencia a fármacos	8 fármacos antituberculosos	Resistencia a los fármacos	1839 isolados de M. tuberculosis	Regresión logística, máquinas de soporte vectorial, bosques aleatorios, modelo de producto de marginales (*product-of-marginals model*) y modelo de mezcla de clase condicional de Bernoulli (*class-conditional Bernoulli mixture model*). Se realizó una comparación entre todos y se escogió aquel con mejor desempeño	El poder predictivo de los modelos depende en gran medida del conjunto de datos utilizados inicialmente para la predicción, por lo que el modelo con mejor desempeño no fue constante cuando se varió el conjunto de datos de partida	AUC > 0,9 para todos los medicamentos bajo el mejor clasificador	>84% para todos los medicamentos bajo el mejor clasificador	>90% para todos los medicamentos bajo el mejor clasificador
Kuo et al., (2017) (160)	Pacientes pediátricos con enfermedad de Kawasaki tratados con inmunoglobulina intravenosa	SNPs	SNPs identificados a partir de un GWAS	Inmunoglobulina intravenosa	Respuesta al tratamiento en términos de reducción de inflamación y presencia o no de lesiones en arterias coronarias	150 pacientes pediátricos	Sistema de puntuación ponderado de riesgo genético (*Weighted Genetic Risk Scoring System*) propuesto por De Jager	El sistema de puntuación pudo discriminar correctamente a respondedores de no respondedores en 2 puntos de corte	AUC = 0,911 para el sistema de puntuación que sólo incluyo variables genéticas y 0,876 para el que incluyó además el género	100%	85,70%
Jenko et al., (2017) (21)	Pacientes adultos con diagnóstico de AR, tratados con	SNPs	Género, presencia de erosiones, factor reumatoideo, anticuerpos anti proteínas	Metotrexato en monoterapia o en combinación	Descontinuación de la terapia debido a eventos adversos	333 pacientes tratados con metotrexato	Regresión penalizada por LASSO	El modelo no pudo predecir correctamente el tiempo hasta descontinuación por eventos adversos ni la descontinuación per se por	AUC = 0,53 para los modelos	NR	NR

Autor, año	Población	Información ómica	Predictores	Intervenciones	Desenlaces	Tamaño de muestra	Técnica de clasificación	Hallazgos	Exactitud en la clasificación	Sensibilidad	Especificidad
	metotrexato por al menos tres meses		citrulinadas, dosis de metotrexato, si está en monoterapia o en combinación y 34 SNPs					eventos adversos, teniendo en cuenta que no todos los pacientes que sufren eventos adversos descontinúan el tratamiento.	farmacogenéticos y clínicos		
Yin et al., (2016) (100)	Pacientes con cáncer de pulmón de células no pequeñas tratados con al menos 2 ciclos de terapias con platino	SNPs	Edad, género, estado de fumador, estadio de la enfermedad, histología y 20 SNPs	Quimioterapias basadas en platino	Respuesta al tratamiento (respuesta completa y respuesta parcial) y toxicidad (eventos adversos)	263 pacientes como cohorte de entrenamiento y 100 pacientes como cohorte de prueba	Naïve Bayes, regresión logística, redes neuronales artificiales, máquinas de soporte vectorial, k vecinos más cercanos, adaboost, agregación de *bootstrap*, árboles de decisión, bosques aleatorios.	El algoritmo Naïve Bayes tuvo el mejor desempeño, comparado con los otros ocho algoritmos. La predicción fue mejor cuando se adicionaron las variables clínicas al modelo. El modelo tuvo mayor poder predictivo en los no fumadores	AUC = 0,80 para respuesta, 0,73 toxicidad general	90% para respuesta y 89% para toxicidad	47% para respuesta y 39% para toxicidad
Kureshi et al., (2016)(109)	Pacientes con cáncer de pulmón de células no pequeñas tratados con inhibidores de tirosina quinasa de primera generación (erlotinib, gefitinib)	Diversos tipos de mutación en el receptor (inserciones, deleciones, duplicaciones, SNPs)	Edad, género, estado de fumador, histología, tipo de ITQ usado, estado del EGFR	Erlotinib o Gefitinib	Respuesta en términos de respuesta completa, parcial, miscelánea y no respuesta como enfermedad estable y enfermedad progresiva	355 pacientes identificados en 34 estudios experimentales y 14 reportes de caso	Máquinas de soporte vectorial, CART, bosques aleatorios, C4.5	El algoritmo con mayor poder predictivo fue SVM, los árboles de decisión tuvieron poderes predictivos comparables, sin embargo, debido a su facil interpretación se prosiguió con su utilización	AUC = 0,76 fue la más alta, obtenida por SVM	NR	NR
González Bosquet et al., (2016) (101)	Pacientes con cáncer de ovario seroso	mRNA (basado en chips)	Expresión diferencial de genes	Quimioterapias basadas en platino	Respuesta al tratamiento (respuesta completa y respuesta incompleta)	450 pacientes	Bosques aleatorios, *Elastic net*, regresión penalizada por LASSO, PAM (*Prediction analysis for microarrays*), Análisis	Bosques aleatorios fue el algoritmo con mejor desempeño consistentemente en los conjuntos de entrenamiento y prueba. La selección inicial de variables identificó un conjunto de 34 genes más relacionados con el desenlace. Los	AUC = 0,74 para el predictor de 34 utilizando bosques aleatorios, 0,68 para el predictor de 422 genes	35% para el predictor de 422 utilizando bosques aleatorios, 13%	90% para el predictor de 422 utilizando bosques aleatorios, 93%

Autor, año	Población	Información ómica	Predictores	Intervenciones	Desenlaces	Tamaño de muestra	Técnica de clasificación	Hallazgos	Exactitud en la clasificación	Sensibilidad	Especificidad
							discriminante diagonal, mínimos cuadrados parciales (PLS), regresión logística penalizada, regresión logística y bosques aleatorios con PLS	modelos realizados con las variables seleccionadas tuvieron mejor desempeño que aquellos que utilizaron los 422 genes iniciales.		para el predictor de 34	para el predictor de 34 genes
Shahid et al., (2016) (156)	Pacientes con cáncer de pulmón de células no pequeñas	mRNA (basado en chips)	Edad, género, estado de fumador, estadio de la enfermedad y expresión diferencial de 8 genes	Quimioterapia adjuvante	Sobrevivencia global	181 pacientes para el conjunto de prueba, 226, 285, 90, 104 y 174 para diferentes validaciones	Regresión multivariada Cox de peligros proporcionales (*multivariate Cox proportional hazard regression*)	La firma de 8 genes utilizada como predictor de respuesta demostró estar significativamente asociada tanto en análisis univariados como multivariados. La validación con 5 conjuntos de datos independientes demostró que la firma de 8 genes clasifica correctamente a los pacientes en alto y bajo riesgo. El criterio de evaluación de la clasificación fue la visualización de las curvas de Kaplan-Meier	NR	93,20%	97,20%
Liang et al., (2016) (203)	Información sobre patologías es inexistente, sólo menciona que son pacientes que presentaron eventos adversos a terapias farmacológicas	SNPs	SNPs en CYP1A2 y CYP2D6 relacionados con eventos adversos a fármacos	No es específico	Eventos adversos hematológicos	83 pacientes	Aprendizaje profundo, redes estocásticas generativas (*Generative Stochastic Networks*)	Para la predicción de cada evento adverso, el algoritmo de aprendizaje profundo desarrollado tuvo menor pérdida de exactitud comparado con LASSO y KNN	LA (Loss of accuracy) entre 11% y 19,8% para GSM	NR	NR

Autor, año	Población	Información ómica	Predictores	Intervenciones	Desenlaces	Tamaño de muestra	Técnica de clasificación	Hallazgos	Exactitud en la clasificación	Sensibilidad	Especificidad
Rizk et al., (2016) (204)	Pacientes con hepatitis C positivos para ARN viral, tratados con peginterferon alfa en combinación con ribavirina	SNPs	Respuesta virológica rápida, alfa-fetoproteína, conteo de plaquetas y 3 SNPs en IL28B, PD-1 y CTLA-4	Tratamiento combinado con peginterferon alfa y ribavirina	Respuesta virológica sostenida (SVR)	200 pacientes	Regresión logística multivariada	No se evaluó el desempeño del modelo en una cohorte de pruena, sin embargo, de acuerdo con los resultados del análisis ROC, el modelo tuvo una capacidad predictiva alta.	AUC = 0,822	71,40%	81,70%
van Dijkhuizen et al., (2015) (25)	Pacientes con artritis reumatoide juvenil (JRA) tratados con metotrexato	SNPs	Categoría de la JLA, evaluación del dolor por parte de paciente, trombocitos, ALT, creatinina y SNPs en proteínas involucradas en el metabolismo del metotrexato	Metotrexato	Intolerancia al metotrexato	152 pacientes	Regresión logística multivariada	El modelo que utilizó únicamente las variables clínicas tuvo mayor poder predictivo que el modelo que utilizó tanto las variables clínicas como las genéticas, por lo que el desarrollo de la puntuación para su utilización en la práctica clínica se basó en el modelo clínico	C-statistic = 66,7% para el modelo clínico y 64,6% para el modelo genético	82% para un puntaje de corte de 6	56,1% para un puntaje de corte de 6
Walter et al., (2015) (157)	Pacientes con leucemia mieloide aguda que recibieron tratamiento con intención curativa	Mutaciones en FLT3 y NPM1	Edad, género, tipo de LMA, citogenética, estatus mutacional	Quimioterapia 7+3 con intención curativa	Respuesta completa y supervivencia libre de recaída a 3, 6 y 12 meses de quimioterapia	4601 pacientes de 4 estudios clínicos	Regresión logística multivariada	Si bien muchos parámetros estaban significativamente asociados con los desenlaces evaluados, su poder predictivo fue mínimo en general. El estatus mutacional de FLT3 y NPM1 fueron los parámetros con mayor poder predictivo	AUC = 0,75, 0,76 y 0,75 para respuesta completa y sobrevivencia libre de recaída en 3, 6 y 12 meses, respectivamente	NR	NR
Kim et al., (2013) (120)	Pacientes con síndrome de déficit de atención e hiperactividad (ADHD) tratados con metilfenidato	SNPs	Variables demográficas (edad, género, peso), clínicas (Escala de puntuación de ADHD - IV, dosis de metilfenidato), medidas neuropsicológicas,	Metilfenidato	Escala de mejoría por impresiones clínicas globales	78 pacientes	Máquinas de soporte vectorial, J48, bosques aleatorios, regresión logística en cresta (*Logistic ridge regression*)	SVM fue el algoritmo que mayor poder predictivo obtuvo. La adición de predictores genéticos a aquellos clínicos utilizados frecuentemente le añadió poder predictivo a los modelos generados en una enfermedad tan heterogénea como el ADHD	AUC = 0,84, 0,61, 0,79, 0,73 para SVM, J48, bosques aleatorios y regresión logística en cresta, respectivamente	NR	NR

Autor, año	Población	Información ómica	Predictores	Intervenciones	Desenlaces	Tamaño de muestra	Técnica de clasificación	Hallazgos	Exactitud en la clasificación	Sensibilidad	Especificidad
			imagenológicas y 4 SNPs en 2 genes								
Kautzky et al., (2015) (205)	Pacientes con desorden de depresión mayor	SNPs	Severidad de los episodios, tendencias suicidas, melancolía, fobia social, ansiedad pánico y 12 SNPs en 5 genes	Tratamiento con SSRIs, inhibidores de la monoamino oxidasa (MAO), antidepresivos tri- y tetra-cíclicos, inhibidores de la recaptación de la noradrenalina (NARI), inhibidores de la recaptación de serotinina-norepinefrina (SNRI) o terapia electroconvulsiva.	Respuesta al tratamiento en la escala *Hamilton Rating Scale for Depression* (HAM-D): > 17 resistente, menor o igual a 17, respondedor	225 pacientes	Bosques aleatorios y k-means	El análisis identificó 4 variables (3 SNPs y una clínica) estrechamente relacionadas con el desenlace en el tratamiento antidepresivo. El poder predictivo de las variables en conjunto se estableció por medio de ORs.	NR	NR	NR
Takahashi et al., (2015) (138)	Pacientes con cáncer gástrico	SNPs	Nivel de creatinina, historial de quimioterapia y 2 SNPs en 2 genes	Tratamiento con fluoropirimidinas (5-FU, S-1)	Respuesta al tratamiento (respuesta parcial y respuesta completa) o no respuesta al tratamiento (no cambio y enfermedad progresiva)	119 pacientes	Regresión logística multivariada	El modelo con el mayor poder predictivo fue aquel que incluyó las 4 variables, de acuerdo con el criterio de la AUC. Sin embargo, utilizando el criterio de la sensibilidad y especificidad, el mejor modelo fue aquel que utilizó únicamente los 2 SNPs	AUC = 0,909	68%	100%
Chen et al., (2014) (139)	Pacientes con cáncer de pulmón de células no pequeñas posterior	mRNA (basado en chips)	Edad, estadio del cáncer y expresión diferencial de 7 genes	No es específico	Sobrevivencia global	2738 pacientes de 17 conjuntos de datos de perfiles de expresión	Regresión multiviariada de supervivencia	El modelo multivariado, cuyo resultado es un índice pronóstico de cáncer de pulmón logró discriminar en categorías de riesgo a los individuos tanto en el	NR	NR	NR

Autor, año	Población	Información ómica	Predictores	Intervenciones	Desenlaces	Tamaño de muestra	Técnica de clasificación	Hallazgos	Exactitud en la clasificación	Sensibilidad	Especificidad
	a cirugía de resección							conjunto de entrenamiento como en el conjunto de prueba			
Kim et al., (2014) (206)	Pacientes con cáncer colorectal tratados con quimioterapia	mRNA (RNA-seq)	Edad, género, estadio del cáncer, localización del tumor, expresión de genes TREM1 y CTGF, y los genes a los que estos regulan.	Quimioterapia adjuvante	Sobrevivencia libre de progresión	566 pacientes en el conjunto de entrenamiento y 229 en el conjunto de prueba	Regresión multivariada Cox	El clasificador logró separar pacientes en alto y bajo riesgo, que podrían beneficiarse con quimioterapia adyuvante	NR	NR	NR
van Asten et al., (2014) (98)	Pacientes con degeneración macular relacionada con la edad	SNPs	Edad, antecedente de diabetes mellitus, agudeza visual antes del tratamiento, SNPs en 3 genes	Ranibizumab	Cambio en la agudeza visual reportada por el paciente a través de un cuestionario	391 pacientes	Regresión logística multivariada	El modelo tuvo un poder predictivo aceptable, aunque el test de bondad de ajuste de este no arrojó resultados significativos (p = 0,69). Junto con el modelo se desarrolló un puntaje predictivo para el riesgo de no respuesta en los pacientes, a partir de los coeficientes de regresión.	AUC = 0,77	NR	NR
KayvanJoo et al., (2014) (110)	Pacientes con hepatitis C positivos para ARN viral, tratados con peginterferon alfa en combinación con ribavirina	Genoma viral	16 atributos de la secuencia de nucleótidos	Tratamiento combinado con peginterferon alfa y ribavirina	Respuesta al tratamiento (inespecífica)	93 muestras virales	Algoritmos de inducción de árboles (*Trees Induction algorithms: Decision Tree,Decision Tree Parallel, Decision Stump and Random Forest*), Máquinas de soporte vectorial, Naïve Bayes y redes neuronales	La mejor predicción fue lograda por Naïve Bayes en las secuencias del subtipo 1b, la mayoría de los atributos seleccionados para la creación de modelos consistieron en dinucleótidos de la secuencia	AUC = 0,59 - 0,94	73% - 93%	50% - 85%

Autor, año	Población	Información ómica	Predictores	Intervenciones	Desenlaces	Tamaño de muestra	Técnica de clasificación	Hallazgos	Exactitud en la clasificación	Sensibilidad	Especificidad
Beerenwinkel et al., (2013) (159)	Pacientes con infección por VIH	Mutaciones en el genoma viral	Edad, género, adherencia al tratamiento, raza, tratamiento farmacológico, SIDA, la barrera genética individualizada (BGI) y la puntuación de susceptibilidad genética (PSG)	Tratamiento con 18 antiretrovirales	Éxito o fallo en la terapia, determinado por el número de copias virales por mL de sangre (50 era el punto de corte)	9231 pacientes	Regresión logística multivariada como algoritmo de predicción final, redes bayesianas conjuntivas isotónicas (*Isotonic Conjunctive Bayesian Network (I-CBN) models*) para la determinación de la barrera genética individualizada	La mejor predicción se dio cuando se utilizaron en conjunto la información clínica, farmacológica, mutacional, BGI y PSG. La inclusión de la BGI, que resume la probabilidad de adquirir mutaciones que confieran resistencia dadas las mutaciones adquiridas previamente aumentó significativamente el poder predictivo de del modelo	AUC = 0,861 para el modelo con BGI, PSG y variables clínicas	NR	NR
Zhang et al., (2013) (207)	Pacientes con cáncer de ovario estadio II-IV del atlas de genoma de cáncer	mRNA (basado en chips)	Perfiles de expresión de genes	Quimioterapia con platino	Sobrevivencia global	No es claro	Net-Cox (network-based Cox proportional hazard model - modelo Cox de peligros proporcionales basado en redes)	Net-cox identifica de manera consistente genes posiblemente relacionados con el desenlace, a través de conjuntos de datos independientes y mejora la predicción de la supervivencia. En general, el poder predictivo aumentaba con respecto al tiempo de seguimiento.	AUC = 0,6 - 0,8 en los diferentes conjuntos de datos.	NR	NR
Jiménez-Sousa et al., (2013) (107)	Pacientes con infección conjunta por hepatitis C y VIH tratados con peginterferon alfa y ribavirina	SNPs	SNPs en IL28RA	Tratamiento combinado con peginterferon alfa y ribavirina	Fallo temprano al tratamiento (Early treatment failure)	291 pacientes	Árbol de decisión	El algoritmo pudo clasificar correctamente el 77,3% de los pacientes en promedio. Los genotipos "favorables" disminuían la tasa de fallo temprano al tratamiento.	AUC = 0,802	NR	NR
Fransen et al., (2012) (23)	Pacientes con artritis reumatoide	SNPs	Género, DAS28 de base, estatus de fumador,	Tratamiento con metotrexato	Respuesta o no respuesta al tratamiento basado en un punto de	75 pacientes	Regresión logística multivariada	El modelo pudo predecir correctamente el 75% de los respondedores. Las variables genéticas no fueron estadísticamente	AUC = 0,77 para el modelo farmacogenético,	NR	NR

Autor, año	Población	Información ómica	Predictores	Intervenciones	Desenlaces	Tamaño de muestra	Técnica de clasificación	Hallazgos	Exactitud en la clasificación	Sensibilidad	Especificidad
			factor reumatoideo, SNPs en 4 genes		corte del DAS28 (2,4)			significativas, la muestra fue bastante pequeña como para establecer relaciones o estructuras más complejas entre variables	0,73 para el modelo no genético		
Vidal-Castiñeira et al., (2012) (208)	Pacientes con infección por el virus de la hepatitis C, en tratamiento de primera línea	SNPs	Carga viral, genotipo del virus, polimorfismos en 2 genes	Tratamiento combinado con peginterferon alfa y ribavirina	Respuesta virológica sostenida (SVR)	407 pacientes	Chi-squared Automatic Interaction Detector (CHAID) algorithm	Los predictores más fuertes en el algoritmo correspondieron a predictores genéticos, aún más que los predictores clínicos como la carga viral o el genotipo del virus	AUC = 0,8145	NR	NR
Neukam et al., (2012) (209)	Pacientes con infección por el virus de la hepatitis C	SNPs	Carga viral, genotipo del virus, polimorfismos en IL28B	Tratamiento combinado con peginterferon alfa y ribavirina	Respuesta virológica sostenida (SVR)	321 pacientes para la cohorte de entrenamiento, 200 para la cohorte de prueba	Árbol de decisión	El desempeño del algoritmo fue similar en las cohortes de entrenamiento y prueba, así como en la población conjunta. De acuerdo con los resultados obtenidos, el algoritmo podría ser usado como herramienta de apoyo en la clínica, para guiar el tratamiento antiviral.	AUC = 0,77 en todos los grupos	86,2%, 87,3% y 84,6% en la población conjunta, cohorte de entrenamiento y pruebe, respectivamente	67,4%, 67% y 68% en la población conjunta, cohorte de entrenamiento y prueba, respectivamente
O'Brien et al., (2011) (163)	Pacientes con infección por el virus de la hepatitis C genotipo 1	SNPs	Carga viral, puntuación de fibrosis y polimorfismo en IL28B	Tratamiento combinado con peginterferon alfa y ribavirina	Respuesta virológica sostenida (SVR)	646 pacientes en la cohorte de entrenamiento y 1121 para la cohorte de prueba	Regresión logística	El modelo farmacogenético y clínico se ajustó mejor a los datos que los modelos únicamente farmacogenético y únicamente clínico. IL28B es un predictor importante en la respuesta al tratamiento con peginterferón y ribavirina	AUC = 0,785, 0,73 y 0,6 para los modelos farmacogenético y clínico; únicamente clínico y únicamente farmacogenético, respectivamente	NR	NR
Alterovitz et al., (2011) (210)	Pacientes con mucositis oral derivada de la radioterapia en el tratamiento del	mRNA (basado en chips)	Expresión diferencial de 51 genes en respondedores, no respondedores y placebo	Tratamiento con gamma-D-glutamil-L-Triptófano	Respuesta de acuerdo con los criterios de la OMS	51 pacientes	Análsis probabilístico de redes bayesianas (Bayesian network probabilistic analysis), junto con	El modelo con el mayor poder predictivo fue aquel desarrollado con la información de expresión diferencial combinado con la información de expresión post-tratamiento. El análisis también identificó clústeres de genes	AUC = 0,99 y 1,00 para el subconjunto de datos mencionado cuando se utilizó el factor de Bayes y el criterio de	NR	NR

Autor, año	Población	Información ómica	Predictores	Intervenciones	Desenlaces	Tamaño de muestra	Técnica de clasificación	Hallazgos	Exactitud en la clasificación	Sensibilidad	Especificidad
	cáncer de cabeza y cuello						análisis funcional de los genes	que podrían estar relacionados con la respuesta al tratamiento.	información akaike como punto de corte, respectivamente		
Kurosaki et al., (2011) (211)	Pacientes con infección por el virus de la hepatitis C genotipo 1b	SNPs	Edad, presencia de fibrosis, LDL, sustitución en el aminoácido 70 del núcleo del virus, polimorfismo en la región ISDR (*IFN sensitivity-determining region* - Región determinante de sensibilidad a interferon)	Tratamiento combinado con peginterferon alfa y ribavirina	Respuesta virológica sostenida (SVR)	304 pacientes para la cohorte de entrenamiento y 201 para la cohorte de prueba	Árbol de decisión	El algoritmo dividió tanto a la cohorte de entrenamiento como a la cohorte de prueba en 6 subgrupos de riesgo con probabilidades diferentes de SVR, que se confirmaron al cruzarlos con la SVR real	NR	NR	NR
Xu et al., (2011) (161)	Pacientes pediátricos con asma moderado a severo	SNPs	Edad, género, VEF1 y tratamiento	Budesonida y nedocromil	Exacerbaciones severas (que requieren visitas a urgencias u hospitalización)	417 pacientes en la cohorte de entrenamiento y 164 en la cohorte de validación	Bosques aleatorios	El algoritmo aumentó su capacidad predictiva a medida que se agregaron SNPs como predictores (comenzando sólo con 4 variables clínicas) hasta alcanzar un punto de meseta en 160 SNPs, cuyo poder predictivo fue significativamente diferente del azar	AUC = 0,66 para el modelo con 160 SNPs (igual resultado con 320 SNPs)	66%	60%
Caocci et al., (2010) (124)	Pacientes con talasemia mayor sometidos a transplante de células hematopoyéticas	SNPs	12 variables clínicas y 12 variables relacionadas con polimorfismos en HLA y KIR	Trasplante de células hematopoyéticas	Reacción injerto versus huésped (Si o No)	68 pacientes en la cohorte de entrenamiento y 10 pacientes en la cohorte de prueba	Regresión logística y redes neuronales artificiales	Las redes neuronales se desempeñaron mejor que la regresión logística en cuanto al poder predictivo	NR	80,5% y 90,1% para la regresión logística y redes neuronales, respectivamente	21,7% y 83,3% para la regresión logística y redes neuronales, respectivamente
Daemen et al., (2009) (119)	Pacientes con cáncer rectal	mRNA (basado en	Genes con baja variación de expresión en todos	Cetuximab, capecitabina y radioterapia	Respuesta histopatológica	40 pacientes	Máquinas de soporte vectorial	El mejor modelo combinó la expresión de 21 genes y 14 proteínas, se	AUC = 0,9870 para el mejor modelo	NR	NR

Autor, año	Población	Información ómica	Predictores	Intervenciones	Desenlaces	Tamaño de muestra	Técnica de clasificación	Hallazgos	Exactitud en la clasificación	Sensibilidad	Especificidad
		chips) y proteómica	los tiempos de toma de muestra					identificó al EGFR como un predictor clave en la respuesta			
Petrovski et al., (2009) (212)	Pacientes con epilepsia	SNPs	5 SNPs seleccionados de 4041 SNPs	Fármacos antiepilépticos	Respuesta farmacológica, definida como ausencia de convulsiones no provocadas	115 pacientes para la cohorte de entrenamiento y dos cohortes de prueba de 63 y 108 pacientes	K-vecinos más próximos	El modelo tuvo un buen poder predictivo en las tres cohortes. La utilización de múltiples SNPs demostró ser mejor predictor que usar asociaciones bivariadas SNP-desenlace	Exactitud: 83,5	89%, 91% y 81% para la cohorte de entrenamiento y las dos cohortes de prueba, respectivamente	69%, 53% y 50% para la cohorte de entrenamiento y las dos cohortes de prueba, respectivamente
Wu et al., (2008) (213)	Pacientes con cáncer de pulmón de células no pequeñas	SNPs	15 SNPs y 8 variables clínicas	Terapia con cisplatino en primera línea	Sobrevivencia global	229 pacientes	Árbol de decisión	El modelo que tuvo en cuenta las variables clínicas, epidemiológicas y genéticas tuvo mayor poder predictivo que el modelo que sólo tuvo en cuenta las variables clínicas. El modelo dividió la población en 3 poblaciones de bajo, medio y alto riesgo, que fueron confirmadas con los datos de sobrevivencia global	AUC = 0,71, 0,73 y 0,83 para los modelos con las variables clínicas, clínicas y epidemiológicas y clínicas, epidemiológicas y farmacogenéticas respectivamente	NR	NR
Lin et al., (2008) (126)	Pacientes con esquizofrenia	SNPs	5 SNPs, edad, género, altura, IMC y peso	Tratamiento con clozapina	Respuesta a la clozapina de acuerdo con la escala de impresión clínica global	69 pacientes en la cohorte de entrenamiento y 24 pacientes en la cohorte de prueba	Redes neuronales artificiales y regresión logística como control	La arquitectura que se ajustó mejor a los datos fue standard *feed-forward, fully-connected, back-propagationneural network* con 25 nodos escondidos. Las redes neuronales tuvieron mayor poder predictivo que la regresión logística	AUC = 0,821 y 0,579 para las redes neuronales y regresión logística, respectivamente	100% y 28,6% para las redes neuronales artificiales y regresión logística, respectivamente	76,5% y 88,2% para las redes neuronales y regresión logística, respectivamente
Modlich et al., (2005) (214)	Pacientes con cáncer de seno positivo para	mRNA (basado en chips)	Expresión diferencial de genes (13 clústeres de	Tratamiento con tamoxifeno	Sobrevivencia libre de metástasis distantes	255 pacientes para la cohorte de entrenamiento y 77, 60, 113 y 112 pacientes en	Combinación de regresiones Cox univariadas	El algoritmo tuvo un bajo poder predictivo en los conjuntos de datos independientes. Sin embargo, los 13 clústeres de genes diferencialmente	NR	NR	NR

Autor, año	Población	Información ómica	Predictores	Intervenciones	Desenlaces	Tamaño de muestra	Técnica de clasificación	Hallazgos	Exactitud en la clasificación	Sensibilidad	Especificidad
	receptor de estrógenos		genes cuya expresión estaba correlacionada)			cuatro cohortes de validación		expresaron probaron ser un predictor importante en el pronóstico del tratamiento			

C. Anexo 3: Descripción del conjunto de datos

VARIABLE	Total (N=219)
EDAD	
Datos faltantes	2
Promedio (DE)	61,613 (12,550)
Rango	28,000 - 93,000
TIEMPO CON AR	
Datos faltantes	2
Promedio (DE)	14,069 (11,202)
Rango	0,000 - 46,000
SEXO	
1	171 (78,1%)
2	48 (21,9%)
ESTRATO	
1	6 (2,7%)
2	64 (29,2%)
3	98 (44,7%)
4	38 (17,4%)
5	9 (4,1%)
6	4 (1,8%)
VIVIENDA	
0	1 (0,5%)
1	175 (79,9%)
2	21 (9,6%)
3	21 (9,6%)

4	1 (0,5%)
ESTADO CIVIL	
1	147 (67,1%)
2	17 (7,8%)
3	24 (11,0%)
4	6 (2,7%)
5	25 (11,4%)
TRABAJO ACTUAL	
0	150 (68,5%)
1	67 (30,6%)
2	2 (0,9%)
PENSIONADO	
0	148 (67,6%)
1	16 (7,3%)
2	54 (24,7%)
3	1 (0,5%)
PRIMARIA COMPLETA	
0	23 (10,5%)
1	195 (89,0%)
3	1 (0,5%)
PRIMARIA	
0	6 (2,7%)
1	1 (0,5%)
2	6 (2,7%)
3	8 (3,7%)
4	3 (1,4%)
5	194 (88,6%)
6	1 (0,5%)
SECUNDARIA COMPLETA	
0	91 (41,6%)
1	127 (58,0%)

3	1 (0,5%)
SECUNDARIA	
0	59 (26,9%)
1	6 (2,7%)
2	3 (1,4%)
3	10 (4,6%)
4	9 (4,1%)
5	5 (2,3%)
6	127 (58,0%)
EDUCACIÓN SUPERIOR COMPLETA	
0	170 (77,6%)
1	49 (22,4%)
EDUCACIÓN SUPERIOR	
0	140 (63,9%)
1	1 (0,5%)
2	13 (5,9%)
3	17 (7,8%)
4	6 (2,7%)
5	37 (16,9%)
6	3 (1,4%)
8	2 (0,9%)
HIPERTENSIÓN ARTERIAL	
0	131 (59,8%)
1	88 (40,2%)
ENFERMEDAD CORONARIA	
0	212 (96,8%)
1	7 (3,2%)
INSUFICIENCIA CARDIACA	
0	217 (99,1%)
1	2 (0,9%)
RVC	

0	218 (99,5%)
1	1 (0,5%)
DIABETES MELLITUS	
0	207 (94,5%)
1	12 (5,5%)
HIPOTIROIDISMO	
0	178 (81,3%)
1	41 (18,7%)
DISLIIDEMIA	
0	204 (93,2%)
1	13 (5,9%)
3	2 (0,9%)
OSTEOPOROSIS	
0	169 (77,2%)
1	50 (22,8%)
SJOGREN	
0	190 (86,8%)
1	29 (13,2%)
LUPUS	
0	214 (97,7%)
1	5 (2,3%)
OTROS ANTECEDENTES	
0	214 (97,7%)
1	5 (2,3%)
EPOC	
0	210 (95,9%)
1	9 (4,1%)
SILICOSIS	
0	219 (100,0%)
TUMORES SÓLIDOS	
0	217 (99,1%)

1	2 (0,9%)
ENFERMEDAD RENAL CRÓNICA	
0	217 (99,1%)
1	2 (0,9%)
FIBROMIALGIA	
0	214 (97,7%)
1	5 (2,3%)
OSTEOARTRITIS	
0	196 (89,5%)
1	23 (10,5%)
REEMPLAZO ARTICULAR	
0	183 (83,6%)
1	15 (6,8%)
2	5 (2,3%)
3	8 (3,7%)
4	8 (3,7%)
OTRAS CIRUGÍAS ORTOPÉDICAS	
Datos faltantes	2
0	177 (81,6%)
1	24 (11,1%)
2	5 (2,3%)
3	6 (2,8%)
4	5 (2,3%)
FRACTURAS	
0	201 (91,8%)
1	17 (7,8%)
4	1 (0,5%)
TABAQUISMO	
0	165 (75,3%)
1	16 (7,3%)
2	38 (17,4%)

ALCOHOL	
0	185 (84,5%)
1	19 (8,7%)
2	15 (6,8%)
SUSTANCIAS PSICOACTIVAS	
0	219 (100,0%)
INFLUENZA	
0	149 (68,0%)
1	70 (32,0%)
NEUMOCOCO	
0	180 (82,2%)
1	39 (17,8%)
HEPATITIS B	
0	211 (96,3%)
1	8 (3,7%)
TUBERCULOSIS	
0	206 (94,1%)
1	13 (5,9%)
HEPATITIS A Y B	
0	214 (97,7%)
1	5 (2,3%)
HEPATITIS C	
0	214 (97,7%)
1	5 (2,3%)
MENARQUIA	
Datos faltantes	4
Promedio (DE)	9,963 (6,204)
Rango	0,000 - 19,000
MENOPAUSIA	
Datos faltantes	8
Promedio (DE)	30,204 (24,007)

Rango	0,000 - 60,000
EMBARAZOS	
Datos faltantes	3
Promedio (DE)	2,227 (2,366)
Rango	0,000 - 13,000
PARTOS	
Datos faltantes	3
Promedio (DE)	1,880 (2,107)
Rango	0,000 - 13,000
PLANIFICACION	
Datos faltantes	1
0	203 (93,1%)
1 (Hormonal)	7 (3,2%)
2 (Barrera)	8 (3,7%)
HOSPITALIZACIÓN POR AR	
0	201 (91,8%)
1	18 (8,2%)
HOSPITALIZACIÓN CARDIOVASCULAR	
0	210 (95,9%)
1	5 (2,3%)
2	2 (0,9%)
3	2 (0,9%)
HOSPITALIZACIÓN POR INFECCIÓN	
0	205 (93,6%)
1	14 (6,4%)
HOSPITALIZACIÓN POR RAM	
0	219 (100,0%)
HOSPITALIZACIÓN POR OTRAS CAUSAS	
0	158 (72,1%)
1	55 (25,1%)
2	2 (0,9%)

4	4 (1,8%)
ALTERACIÓN DE LA TERAPIA FARMACOLÓGICA	
0	156 (71,2%)
1	63 (28,8%)
TPO_MDALTER	
0	156 (71,2%)
1	5 (2,3%)
2	7 (3,2%)
3	31 (14,2%)
4	20 (9,1%)
MDALTER_ACTUAL	
0	213 (97,3%)
1	5 (2,3%)
2	1 (0,5%)
ANTECEDENTE FAMILIAR DE AR	
0	148 (67,6%)
1	57 (26,0%)
2	14 (6,4%)
ANTECEDENTE FAMILIAR DE LES	
0	202 (92,2%)
1	11 (5,0%)
2	6 (2,7%)
ANTECEDENTE FAMILIAR DE SJORGEN	
0	216 (98,6%)
1	2 (0,9%)
2	1 (0,5%)
ANTECEDENTE FAMILIAR DE HIPERTIROIDISMO	
0	210 (95,9%)
1	9 (4,1%)

OTROS ANTECEDENTES	
Datos faltantes	25
0	180 (92,8%)
1	14 (7,2%)
METOTREXATE	
1	155 (70,8%)
2	14 (6,4%)
3	49 (22,4%)
4	1 (0,5%)
CONSUMO CONCOMITANTE DE MEDICAMENTOS DE SÍNTESIS CON MTX	
0	106 (48,4%)
1	113 (51,6%)
CONSUMO CONCOMITANTE DE MEDICAMENTOS BIOLÓGICOS CON MTX	
0	185 (84,5%)
1	34 (15,5%)
LEFLUNOMIDA	
0	106 (48,4%)
1	75 (34,2%)
2	13 (5,9%)
3	17 (7,8%)
4	8 (3,7%)
SULFASALAZINA	
0	132 (60,3%)
1	31 (14,2%)
2	22 (10,0%)
3	15 (6,8%)
4	19 (8,7%)
ANTIMALARCO	
0	93 (42,5%)

1	41 (18,7%)
2	21 (9,6%)
3	54 (24,7%)
4	10 (4,6%)
AZATIOPRINA	
0	216 (98,6%)
2	3 (1,4%)
ESTEROIDES	
0	49 (22,4%)
1	123 (56,2%)
2	9 (4,1%)
3	8 (3,7%)
4	30 (13,7%)
ETANERCEPT	
0	193 (88,1%)
1	17 (7,8%)
2	5 (2,3%)
3	3 (1,4%)
4	1 (0,5%)
ADALIMUMAB	
0	191 (87,2%)
1	12 (5,5%)
2	13 (5,9%)
3	3 (1,4%)
INFLIXIMAB	
0	200 (91,3%)
1	3 (1,4%)
2	9 (4,1%)
3	7 (3,2%)
CERTOZUMAB	
0	213 (97,3%)

1	3 (1,4%)
2	1 (0,5%)
3	2 (0,9%)
GOLIMUMAB	
0	214 (97,7%)
1	4 (1,8%)
2	1 (0,5%)
ABATACEPT	
0	201 (91,8%)
1	15 (6,8%)
2	3 (1,4%)
TOCILIZUMAB	
Datos faltantes	1
0	214 (98,2%)
1	2 (0,9%)
3	1 (0,5%)
4	1 (0,5%)
RITUXIMAB	
0	200 (91,3%)
1	14 (6,4%)
2	2 (0,9%)
3	2 (0,9%)
4	1 (0,5%)
TOFACITINIB	
0	216 (98,6%)
1	3 (1,4%)
OJO SECO	
0	137 (62,6%)
1	12 (5,5%)
2	29 (13,2%)
3	41 (18,7%)

BOCA SECA	
0	133 (60,7%)
1	9 (4,1%)
2	29 (13,2%)
3	48 (21,9%)
SEQUEDAD GENITAL	
0	132 (85,2%)
1	2 (1,3%)
2	10 (6,5%)
3	11 (7,1%)
CLASIFICACIÓN FUNCIONAL	
1	60 (38,7%)
2	62 (40,0%)
3	20 (12,9%)
4	13 (8,4%)
DAS28 EN 4 CATEGORÍAS	
Datos faltantes	1
1	54 (24,8%)
2	37 (17,0%)
3	95 (43,6%)
4	32 (14,7%)
IMC	
Datos faltantes	100
Promedio (DE)	26,172 (4,964)
Rango	3,420 - 41,000
rs1045642	
AA	54 (24,7%)
AG	104 (47,5%)
GG	61 (27,9%)
rs1051266	
CC	70 (32,0%)

CT	113 (51,6%)
TT	36 (16,4%)
rs1061622	
Datos faltantes	1
GG	4 (1,8%)
GT	44 (20,2%)
TT	170 (78,0%)
rs1061631	
Datos faltantes	1
AA	2 (0,9%)
AG	38 (17,4%)
GG	178 (81,7%)
rs10919563	
AA	27 (12,3%)
AG	104 (47,5%)
GG	88 (40,2%)
rs11052877	
AA	100 (45,7%)
AG	97 (44,3%)
GG	22 (10,0%)
rs11545078	
AG	24 (11,0%)
GG	195 (89,0%)
rs12083537	
AA	168 (76,7%)
AG	47 (21,5%)
GG	4 (1,8%)
rs1799724	
CC	120 (54,8%)
CT	93 (42,5%)
TT	6 (2,7%)

rs1799964	
CC	13 (5,9%)
CT	84 (38,4%)
TT	122 (55,7%)
rs1800795	
CC	12 (5,5%)
CG	71 (32,4%)
GG	136 (62,1%)
rs1800896	
CC	20 (9,1%)
CT	88 (40,2%)
TT	111 (50,7%)
rs1801131	
GG	5 (2,3%)
GT	49 (22,4%)
TT	165 (75,3%)
rs1801133	
AA	56 (25,6%)
AG	107 (48,9%)
GG	56 (25,6%)
rs1801157	
CC	143 (65,3%)
CT	71 (32,4%)
TT	5 (2,3%)
rs1801274	
AA	48 (21,9%)
AG	95 (43,4%)
GG	76 (34,7%)
rs1883112	
AA	81 (37,0%)
AG	107 (48,9%)

GG	31 (14,2%)
rs20575	
Datos faltantes	2
CC	30 (13,8%)
CG	97 (44,7%)
GG	90 (41,5%)
rs2229109	
CC	209 (95,4%)
CT	9 (4,1%)
TT	1 (0,5%)
rs2372536	
CC	112 (51,1%)
CG	93 (42,5%)
GG	14 (6,4%)
rs3397	
CC	21 (9,6%)
CT	100 (45,7%)
TT	98 (44,7%)
rs361525	
AA	1 (0,5%)
AG	25 (11,4%)
GG	193 (88,1%)
rs3761847	
Datos faltantes	1
AA	105 (48,2%)
AG	89 (40,8%)
GG	24 (11,0%)
rs3794271	
AA	35 (16,0%)
AG	110 (50,2%)
GG	74 (33,8%)

rs3821353	
Datos faltantes	1
GG	108 (49,5%)
GT	64 (29,4%)
TT	46 (21,1%)
rs3849942	
CC	163 (74,4%)
CT	55 (25,1%)
TT	1 (0,5%)
rs394581	
CC	9 (4,1%)
CT	46 (21,0%)
TT	164 (74,9%)
rs396991	
AA	128 (58,4%)
AC	80 (36,5%)
CC	11 (5,0%)
rs4148396	
CC	118 (53,9%)
CT	84 (38,4%)
TT	17 (7,8%)
rs4329505	
CT	56 (25,6%)
TT	163 (74,4%)
rs437943	
CC	43 (19,6%)
CT	100 (45,7%)
TT	76 (34,7%)
rs4411591	
CC	191 (87,2%)
CT	26 (11,9%)

TT	2 (0,9%)
rs4451422	
Datos faltantes	1
AA	71 (32,6%)
AC	110 (50,5%)
CC	37 (17,0%)
rs4673993	
CC	17 (7,8%)
CT	93 (42,5%)
TT	109 (49,8%)
rs4750316	
Datos faltantes	2
CC	8 (3,7%)
CG	38 (17,5%)
GG	171 (78,8%)
rs4846051	
Datos faltantes	1
AA	208 (95,4%)
AG	10 (4,6%)
rs4982133	
AA	9 (4,1%)
AC	38 (17,4%)
CC	172 (78,5%)
rs4986790	
AA	206 (94,1%)
AG	13 (5,9%)
rs548234	
CC	9 (4,1%)
CT	76 (34,7%)
TT	134 (61,2%)
rs5751876	

CC	70 (32,0%)
CT	99 (45,2%)
TT	50 (22,8%)
rs5760410	
AA	63 (28,8%)
AG	101 (46,1%)
GG	55 (25,1%)
rs6028945	
GG	179 (81,7%)
GT	39 (17,8%)
TT	1 (0,5%)
rs6064463	
CC	60 (27,4%)
CT	117 (53,4%)
TT	42 (19,2%)
rs6071980	
CC	13 (5,9%)
CT	96 (43,8%)
TT	110 (50,2%)
rs6138150	
CC	23 (10,5%)
CT	86 (39,3%)
TT	110 (50,2%)
rs6427528	
AA	20 (9,1%)
AG	73 (33,3%)
GG	126 (57,5%)
rs6506569	
CC	83 (37,9%)
CT	98 (44,7%)
TT	38 (17,4%)

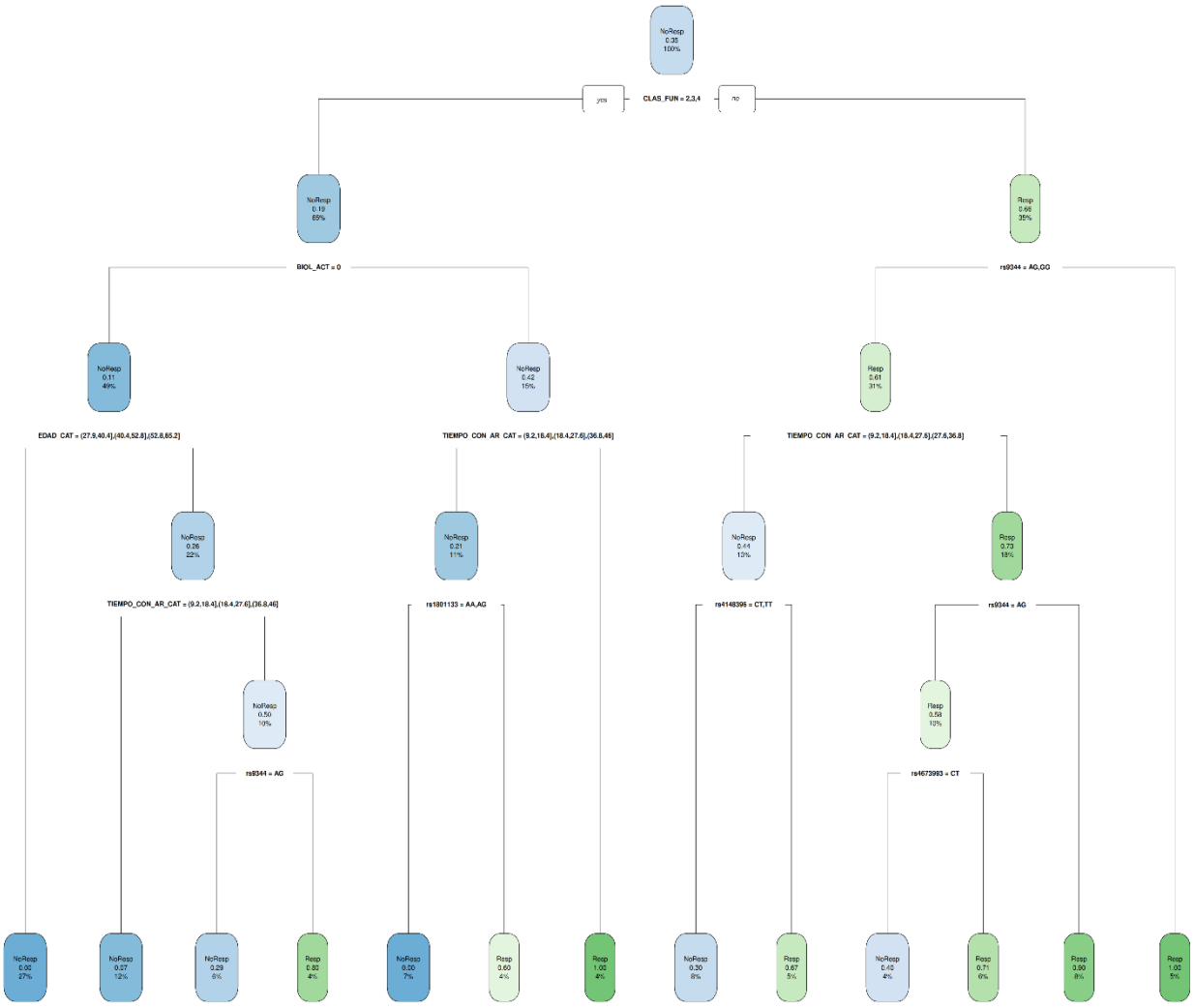
rs6691117	
AA	73 (33,3%)
AG	109 (49,8%)
GG	37 (16,9%)
rs6822844	
GG	195 (89,0%)
GT	24 (11,0%)
rs6920220	
AA	1 (0,5%)
AG	50 (22,8%)
GG	168 (76,7%)
rs7046653	
AA	25 (11,4%)
AG	114 (52,1%)
GG	80 (36,5%)
rs70991108	
DD	71 (32,4%)
ID	95 (43,4%)
II	53 (24,2%)
rs719235	
AA	7 (3,2%)
AC	62 (28,3%)
CC	150 (68,5%)
rs7279445	
CC	33 (15,1%)
CT	123 (56,2%)
TT	63 (28,8%)
rs7499	
AA	35 (16,0%)
AG	112 (51,1%)
GG	72 (32,9%)

rs7527798	
CC	7 (3,2%)
CT	46 (21,0%)
TT	166 (75,8%)
rs7563206	
CC	84 (38,4%)
CT	96 (43,8%)
TT	39 (17,8%)
rs7574865	
GG	73 (33,3%)
GT	107 (48,9%)
TT	39 (17,8%)
rs7624766	
AA	49 (22,4%)
AG	110 (50,2%)
GG	60 (27,4%)
rs774359	
CC	1 (0,5%)
CT	59 (26,9%)
TT	159 (72,6%)
rs854547	
Datos faltantes	1
AA	62 (28,4%)
AG	115 (52,8%)
GG	41 (18,8%)
rs854548	
AA	25 (11,4%)
AG	87 (39,7%)
GG	107 (48,9%)
rs854555	
AA	37 (16,9%)

AC	115 (52,5%)
CC	67 (30,6%)
rs868856	
AA	25 (11,4%)
AG	114 (52,1%)
GG	80 (36,5%)
rs928655	
AA	85 (38,8%)
AG	111 (50,7%)
GG	23 (10,5%)
rs9344	
AA	35 (16,0%)
AG	99 (45,2%)
GG	85 (38,8%)
rs9514828	
CC	122 (55,7%)
CT	77 (35,2%)
TT	20 (9,1%)
rs983332	
GG	158 (72,1%)
GT	56 (25,6%)
TT	5 (2,3%)
rs9977268	
Datos faltantes	1
CC	119 (54,6%)
CT	87 (39,9%)
TT	12 (5,5%)
EDAD_CAT	
Datos faltantes	2
(27,9,41]	14 (6,5%)
(41,54]	45 (20,7%)

(54,67]	85 (39,2%)
(67,80]	65 (30,0%)
(80,93,1]	8 (3,7%)
TIEMPO CON AR CATEGÓRICA	
Datos faltantes	2
(-0,046,9,2]	93 (42,9%)
(9,2,18,4]	56 (25,8%)
(18,4,27,6]	39 (18,0%)
(27,6,36,8]	18 (8,3%)
(36,8,46]	11 (5,1%)
DAS28 EN 2 CATEGORÍAS	
Datos faltantes	4
No Respondedor	148 (68,8%)
Respondedor	67 (31,2%)

D. Anexo 4: Árbol de decisión obtenido con el conjunto de datos final



10. Referencias

1. McInnes IB, Schett G. The pathogenesis of rheumatoid arthritis. N Engl J Med. 2011.365(23):2205–19.

2. Maradit-Kremers H, Nicola PJ, Crowson CS, O'Fallon WM, Gabriel SE. Patient, disease, and therapy-related factors that influence discontinuation of disease-modifying antirheumatic drugs: A population-based incidence cohort of patients with rheumatoid arthritis. J Rheumatol. 2006.33(2):248–55.

3. Singh JA, Saag KG, Bridges SL, Akl EA, Bannuru RR, Sullivan MC, et al. 2015 American College of Rheumatology Guideline for the Treatment of Rheumatoid Arthritis. Arthritis Care Res (Hoboken). 2016.68(1):1–25.

4. Smolen JS, Landewé R, Bijlsma J, Burmester G, Chatzidionysiou K, Dougados M, et al. EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2016 update. Ann Rheum Dis. 2017.76(6):960–77.

5. Boonen A, Severens JL. The burden of illness of rheumatoid arthritis. Clin Rheumatol. 2011.30(SUPPL. 1):3–8.

6. Umićević Mirkov M, Coenen MJ. Pharmacogenetics of disease-modifying antirheumatic drugs in rheumatoid arthritis: towards personalized medicine. Pharmacogenomics. 2013.14(4):425–44.

7. Padmanabhan S. Handbook of Pharmacogenomics and Stratified Medicine. Padmanabhan S, editor. 2014. 1–1119 p.

8. Beyer K, Zaura E, Brandt BW, Buijs MJ, Brun JG, Crielaard W, et al. Subgingival microbiome of rheumatoid arthritis patients in relation to their disease status and

periodontal health. PLoS One. 2018.13(9):e0202278.

9. Visser B, Brinkman I, Van De Laar MA. Personalized medicine in rheumatoid arthritis: Rationale and clinical evidence. Clin Investig (Lond). 2012.2(8):797–802.

10. Littman BH. Translational strategies to implement personalized medicine: rheumatoid arthritis examples. Per Med. 2009.6(4):429–37.

11. Mathé E, Hays J, Stover D, Chen J. The Omics Revolution Continues: The Maturation of High-Throughput Biological Data Sources. Yearb Med Inform. 2018.27(01):211–22.

12. Walsh AM, Whitaker JW, Huang CC, Cherkas Y, Lamberth SL, Brodmerkel C, et al. Integrative genomic deconvolution of rheumatoid arthritis GWAS loci into gene and cell type associations. Genome Biol. 2016.17(1):79.

13. Maldonado-Montoro M, Canadas-Garre M, Gonzalez-Utrilla A, Plaza-Plaza JC, Calleja-Hernandez MY. Genetic and clinical biomarkers of tocilizumab response in patients with rheumatoid arthritis. Pharmacol Res. 2016.111:264–71.

14. Moya P, Salazar J, Arranz MJ, Díaz-Torné C, del Río E, Casademont J, et al. Methotrexate pharmacokinetic genetic variants are associated with outcome in rheumatoid arthritis patients. Pharmacogenomics. 2016.17(1):25–9.

15. Hider SL, Thomson W, Mack LF, Armstrong DJ, Shadforth M, Bruce IN. Polymorphisms within the adenosine receptor 2a gene are associated with adverse events in RA patients treated with MTX. Rheumatology (Oxford). 2008.47(8):1156–9.

16. Chandran V, Siannis F, Rahman P, Pellett FJ, Farewell VT, Gladman DD. Folate pathway enzyme gene polymorphisms and the efficacy and toxicity of methotrexate in psoriatic arthritis. J Rheumatol. 2010.37(7):1508–12.

17. Potter C, Cordell HJ, Barton A, Daly AK, Hyrich KL, Mann DA, et al. Association between anti-tumour necrosis factor treatment response and genetic variants within the TLR and NF$κ$B signalling pathways. Ann Rheum Dis. 2010.69(7):1315–20.

18. Netz U, Carter JV, Eichenberger MR, Dryden GW, Pan J, Rai SN, et al. Genetic polymorphisms predict response to anti-tumor necrosis factor treatment in Crohn’s

disease. WORLD J Gastroenterol. 2017.23(27):4958–67.

19. Lech-Maranda E, Grzybowska-Izydorczyk O, Wyka K, Mlynarski W, Borowiec M, Antosik K, et al. Serum tumor necrosis factor-alpha and interleukin-10 levels as markers to predict outcome of patients with chronic lymphocytic leukemia in different risk groups defined by the IGHV mutation status. Arch Immunol Ther Exp (Warsz). 2012.60(6):477–86.

20. Liu C, Batliwalla F, Li W, Lee A, Roubenoff R, Beckman E, et al. Genome-wide association scan identifies candidate polymorphisms associated with differential response to anti-TNF treatment in rheumatoid arthritis. Mol Med. 14(9–10):575–81.

21. Jenko B, Lusa L, Tomsic M, Praprotnik S, Dolzan V. Clinical-pharmacogenetic predictive models for MTX discontinuation due to adverse events in rheumatoid arthritis. Pharmacogenomics J. 2017.17(5):412–8.

22. Jenko B, Tomšič M, Jekić B, Milić V, Dolžan V, Praprotnik S. Clinical Pharmacogenetic Models of Treatment Response to Methotrexate Monotherapy in Slovenian and Serbian Rheumatoid Arthritis Patients: Differences in Patient's Management May Preclude Generalization of the Models. Front Pharmacol. 2018.9:20.

23. Fransen J, Kooloos WM, Wessels JAM, Huizinga TWJ, Guchelaar H-J, van Riel PLCM, et al. Clinical pharmacogenetic model to predict response of MTX monotherapy in patients with established rheumatoid arthritis after DMARD failure. Pharmacogenomics. 2012.13(9):1087–94.

24. de Rotte MCFJ, Pluijm SMF, de Jong PHP, Bulatović Ćalasan M, Wulffraat NM, Weel AEAM, et al. Development and validation of a prognostic multivariable model to predict insufficient clinical response to methotrexate in rheumatoid arthritis. Abu-Shakra M, editor. PLoS One. 2018.13(12):e0208534.

25. van Dijkhuizen EHP, Bulatovic Calasan M, Pluijm SMF, de Rotte MCFJ, Vastert SJ, Kamphuis S, et al. Prediction of methotrexate intolerance in juvenile idiopathic arthritis: a prospective, observational cohort study. Pediatr Rheumatol Online J.

2015.13:5.

26. Marwa OS, Kalthoum T, Wajih K, Kamel H. Association of IL17A and IL17F genes with rheumatoid arthritis disease and the impact of genetic polymorphisms on response to treatment. Immunol Lett. 2017.183:24–36.

27. Boughrara W, Benzaoui A, Aberkane M, Moghtit FZ, Dorgham S, Lardjam-Hetraf AS, et al. No correlation between MTHFR c.677 C >T, MTHFR c.1298 A >C, and ABCB1 c.3435 C >T polymorphisms and methotrexate therapeutic outcome of rheumatoid arthritis in West Algerian population. Inflamm Res. 2017.66(6):505–13.

28. Soukup T, Dosedel M, Pavek P, Nekvindova J, Barvik I, Bubancova I, et al. The impact of C677T and A1298C MTHFR polymorphisms on methotrexate therapeutic response in East Bohemian region rheumatoid arthritis patients. Rheumatol Int. 2015.35(7):1149–61.

29. Lima A, Seabra V, Bernardes M, Azevedo R, Sousa H, Medeiros R. Role of key TYMS polymorphisms on methotrexate therapeutic outcome in portuguese rheumatoid arthritis patients. PLoS One. 2014.9(10):e108165.

30. Lee YH, Ji JD, Bae S-C, Song GG. Associations between tumor necrosis factor-alpha (TNF-alpha) -308 and -238 G/A polymorphisms and shared epitope status and responsiveness to TNF-alpha blockers in rheumatoid arthritis: a metaanalysis update. J Rheumatol. 2010.37(4):740–6.

31. Cuchacovich M, Soto L, Edwardes M, Gutierrez M, Llanos C, Pacheco D, et al. Tumour necrosis factor (TNF)alpha -308 G/G promoter polymorphism and TNFalpha levels correlate with a better response to adalimumab in patients with rheumatoid arthritis. Scand J Rheumatol. 2006.35(6):435–40.

32. Croia C, Bursi R, Sutera D, Petrelli F, Alunno A, Puxeddu I. One year in review 2019: pathogenesis of rheumatoid arthritis. Clin Exp Rheumatol. 2019.37(3):347–57.

33. Smolen JS, Aletaha D, Barton A, Burmester GR, Emery P, Firestein GS, et al. Rheumatoid arthritis. Nat Rev Dis Prim. 2018.4(1):18001.

34. Aletaha D, Smolen JS. Diagnosis and Management of Rheumatoid Arthritis. JAMA.

2018.320(13):1360.

35. Tan EM, Smolen JS. Historical observations contributing insights on etiopathogenesis of rheumatoid arthritis and role of rheumatoid factor. J Exp Med. 2016.213(10):1937–50.

36. Redlich K, Smolen JS. Inflammatory bone loss: pathogenesis and therapeutic intervention. Nat Rev Drug Discov. 2012.11(3):234–50.

37. Aletaha D, Smolen JS. Joint damage in rheumatoid arthritis progresses in remission according to the Disease Activity Score in 28 joints and is driven by residual swollen joints. Arthritis Rheum. 2011.63(12):3702–11.

38. Wallach D. The cybernetics of TNF: Old views and newer ones. Semin Cell Dev Biol. 2016.50:105–14.

39. Aggarwal R, Ringold S, Khanna D, Neogi T, Johnson SR, Miller A, et al. Distinctions between diagnostic and classification criteria? Arthritis Care Res (Hoboken). 2015.67(7):891–7.

40. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO, et al. 2010 Rheumatoid arthritis classification criteria: An American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis Rheum. 2010.62(9):2569–81.

41. Díaz-Rojas JA, Dávila-Ramírez FA, Quintana-López G, Aristizábal-Gutiérrez F, Brown P. Prevalencia de artritis reumatoide en Colombia: una aproximación basada en la carga de la enfermedad durante el año 2005. Rev Colomb Reumatol. 2016.23(1):11–6.

42. Fernández-Ávila DG, Rincón-Riaño DN, Bernal-Macías S, Gutiérrez Dávila JM, Rosselli D. Prevalencia de la artritis reumatoide en Colombia según información del Sistema Integral de Información de la Protección Social. Rev Colomb Reumatol. 2019.26(2):83–7.

43. Cuenta de Alto Costo. Situación de la artritis reumatoide en Colombia 2016

[Internet]. Bogotá D.C.: Fondo Colombiano de Enfermedades de Alto Costo; 2017. p. 1–81.

44. Eriksson JK, Johansson K, Askling J, Neovius M. Costs for hospital care, drugs and lost work days in incident and prevalent rheumatoid arthritis: how large, and how are they distributed? Ann Rheum Dis. 2015.74(4):648–54.

45. Sokka T, Kautiainen H, Pincus T, Verstappen SMM, Aggarwal A, Alten R, et al. Work disability remains a major problem in rheumatoid arthritis in the 2000s: data from 32 countries in the QUEST-RA study. Arthritis Res Ther. 2010.12(2):R42.

46. Machado J, Moncada JC, Pineda R. Perfil de utilización de los anti-factor de necrosis tumoral en pacientes de Colombia. Biomédica. 2011.31(2):250.

47. Quintana G, Mora C, González A, Díaz JD. Costos directos de la artritis reumatoide temprana en el primer año de atención: simulación de tres situaciones clínicas en un hospital universitario de tercer nivel en Colombia. Biomédica. 2009.29(1):43.

48. Montoya N, Gómez L, Vélez M, Rosselli D. Costos directos del tratamiento de pacientes con artritis reumatoide en Medellín, Colombia. Rev Colomb Reumatol. 2011.18(1):26–33.

49. Aletaha D, Ward MM, Machold KP, Nell VPK, Stamm T, Smolen JS. Remission and active disease in rheumatoid arthritis: defining criteria for disease activity states. Arthritis Rheum. 2005.52(9):2625–36.

50. Mierau M, Schoels M, Gonda G, Fuchs J, Aletaha D, Smolen JS. Assessing remission in clinical practice. Rheumatology (Oxford). 2007.46(6):975–9.

51. Klarenbeek NB, Guler-Yuksel M, van der Kooij SM, Han KH, Ronday HK, Kerstens PJSM, et al. The impact of four dynamic, goal-steered treatment strategies on the 5-year outcomes of rheumatoid arthritis patients in the BeSt study. Ann Rheum Dis. 2011.70(6):1039–46.

52. Aletaha D, Smolen J, Ward MM. Measuring function in rheumatoid arthritis: Identifying reversible and irreversible components. Arthritis Rheum. 2006.54(9):2784–92.

53. van der Heijde DM, van ’t Hof M, van Riel PL, van de Putte LB. Development of a disease activity score based on judgment in clinical practice by rheumatologists. J Rheumatol. 1993.20(3):579–81.

54. Prevoo ML, van ’t Hof MA, Kuper HH, van Leeuwen MA, van de Putte LB, van Riel PL. Modified disease activity scores that include twenty-eight-joint counts. Development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. Arthritis Rheum. 1995.38(1):44–8.

55. Bakker MF, Jacobs JWG, Verstappen SMM, Bijlsma JWJ. Tight control in the treatment of rheumatoid arthritis: efficacy and feasibility. Ann Rheum Dis. 2007.66 Suppl 3:iii56-60.

56. Fransen J, Stucki G, van Riel PLCM. Rheumatoid arthritis measures: Disease Activity Score (DAS), Disease Activity Score-28 (DAS28), Rapid Assessment of Disease Activity in Rheumatology (RADAR), and Rheumatoid Arthritis Disease Activity Index (RADAI). Arthritis Rheum. 2003.49(S5):S214–24.

57. Smolen JS, van der Heijde D, Machold KP, Aletaha D, Landewé R. Proposal for a new nomenclature of disease-modifying antirheumatic drugs: Table 1. Ann Rheum Dis. 2014.73(1):3–5.

58. Hoffmeister RT. Methotrexate therapy in rheumatoid arthritis: 15 years experience. Am J Med. 1983.75(6):69–73.

59. Grosflam J, Weinblatt ME. Methotrexate: mechanism of action, pharmacokinetics, clinical indications, and toxicity. Curr Opin Rheumatol. 1991.3(3):363–8.

60. Cronstein BN. The mechanism of action of methotrexate. Rheum Dis Clin North Am. 1997.23(4):739–55.

61. Segal R, Tartakovsky B, Rafael B. Methotrexate: Mechanism of Action in Rheumatoid Arthriti. Semin Arthritis Rheum. 1990.20(3):190–9.

62. Ministerio de Salud y Protección Social. Departamento Administrativo de Ciencia Tecnología e Innovación. Guía de Práctica Clínica para la detección temprana,

diagnóstico y tratamiento de la artritis reumatoide. Guía No. GPC-2014-26. 2014; p. 1–876.

63. Kiely P, Walsh D, Williams R, Young A. Outcome in rheumatoid arthritis patients with continued conventional therapy for moderate disease activity--the early RA network (ERAN). Rheumatology. 2011.50(5):926–31.

64. Porter D, van Melckebeke J, Dale J, Messow CM, McConnachie A, Walker A, et al. Tumour necrosis factor inhibition versus rituximab for patients with rheumatoid arthritis who require biological treatment (ORBIT): an open-label, randomised controlled, non-inferiority, trial. Lancet (London, England). 2016.388(10041):239–47.

65. Weinblatt ME, Schiff M, Valente R, van der Heijde D, Citera G, Zhao C, et al. Head-to-head comparison of subcutaneous abatacept versus adalimumab for rheumatoid arthritis: findings of a phase IIIb, multinational, prospective, randomized study. Arthritis Rheum. 2013.65(1):28–38.

66. Fleischmann R, Mysler E, Hall S, Kivitz AJ, Moots RJ, Luo Z, et al. Efficacy and safety of tofacitinib monotherapy, tofacitinib with methotrexate, and adalimumab with methotrexate in patients with rheumatoid arthritis (ORAL Strategy): a phase 3b/4, double-blind, head-to-head, randomised controlled trial. Lancet (London, England). 2017.390(10093):457–68.

67. Fleischmann R, Schiff M, van der Heijde D, Ramos-Remus C, Spindler A, Stanislav M, et al. Baricitinib, Methotrexate, or Combination in Patients With Rheumatoid Arthritis and No or Limited Prior Disease-Modifying Antirheumatic Drug Treatment. Arthritis Rheumatol (Hoboken, NJ). 2017.69(3):506–17.

68. Nam JL, Takase-Minegishi K, Ramiro S, Chatzidionysiou K, Smolen JS, van der Heijde D, et al. Efficacy of biological disease-modifying antirheumatic drugs: a systematic literature review informing the 2016 update of the EULAR recommendations for the management of rheumatoid arthritis. Ann Rheum Dis. 2017.76(6):1113–36.

69. Bang LM, Keating GM. Adalimumab - A Review of its Use in Rheumatoid Arthritis. BioDrugs. 2004.18(2):121–39.

70. Combe B, Logeart I, Belkacemi MC, Dadoun S, Schaeverbeke T, Daurès JP, et al. Comparison of the long-term outcome for patients with rheumatoid arthritis with persistent moderate disease activity or disease remission during the first year after diagnosis: data from the ESPOIR cohort. Ann Rheum Dis. 2015.74(4):724–9.

71. Jayakumar K, Norton S, Dixey J, James D, Gough A, Williams P, et al. Sustained clinical remission in rheumatoid arthritis: prevalence and prognostic factors in an inception cohort of patients treated with conventional DMARDS. Rheumatology (Oxford). 2012.51(1):169–75.

72. Svensson B, Andersson MLE, Bala S-V, Forslind K, Hafström I, BARFOT study group. Long-term sustained remission in a cohort study of patients with rheumatoid arthritis: choice of remission criteria. BMJ Open. 2013.3(9):e003554.

73. Owen SA, Lunt M, Hider SL, Bruce IN, Barton A, Thomson W. Testing pharmacogenetic indices to predict efficacy and toxicity of methotrexate monotherapy in a rheumatoid arthritis patient cohort. Arthritis Rheum. 2010.62(12):3827–9.

74. Zhu H, Deng F-Y, Mo X-B, Qiu Y-H, Lei S-F. Pharmacogenetics and pharmacogenomics for rheumatoid arthritis responsiveness to methotrexate treatment: the 2013 update. Pharmacogenomics. 2014.15(4):551–66.

75. Malik F, Ranganathan P. Methotrexate pharmacogenetics in rheumatoid arthritis: a status report. Pharmacogenomics. 2013.14(3):305–14.

76. Cuchacovich M, Ferreira L, Aliste M, Soto L, Cuenca J, Cruzat A, et al. Tumour necrosis factor-alpha (TNF-alpha) levels and influence of -308 TNF-alpha promoter polymorphism on the responsiveness to infliximab in patients with rheumatoid arthritis. Scand J Rheumatol. 2004.33(4):228–32.

77. O'Rielly DD, Roslin NM, Beyene J, Pope A, Rahman P. TNF-alpha-308 G/A polymorphism and responsiveness to TNF-alpha blockade therapy in moderate to severe rheumatoid arthritis: a systematic review and meta-analysis. Pharmacogenomics J. 2009.9(3):161–7.

78. Umicevic Mirkov M, Cui J, Vermeulen SH, Stahl EA, Toonen EJM, Makkinje RR, et al. Genome-wide association analysis of anti-TNF drug response in patients with rheumatoid arthritis. Ann Rheum Dis. 2013.72(8):1375–81.

79. Montes A, Perez-Pampin E, Narváez J, Cañete JD, Navarro-Sarabia F, Moreira V, et al. Association of FCGR2A with the response to infliximab treatment of patients with rheumatoid arthritis. Pharmacogenet Genomics. 2014.24(5):238–45.

80. Lee YH, Bae S-C. Associations between PTPRC rs10919563 A/G and FCGR2A R131H polymorphisms and responsiveness to TNF blockers in rheumatoid arthritis: a meta-analysis. Rheumatol Int. 2016.36(6):837–44.

81. Cui J, Saevarsdottir S, Thomson B, Padyukov L, van der Helm-van Mil AHM, Nititham J, et al. Rheumatoid arthritis risk allele PTPRC is also associated with response to anti-tumor necrosis factor alpha therapy. Arthritis Rheum. 2010.62(7):1849–61.

82. Ferreiro-Iglesias A, Montes A, Perez-Pampin E, Canete JD, Raya E, Magro-Checa C, et al. Replication of PTPRC as genetic biomarker of response to TNF inhibitors in patients with rheumatoid arthritis. Pharmacogenomics J. 2016.16(2):137–40.

83. Dávila-Fajardo CL, van der Straaten T, Baak-Pablo R, Medarde Caballero C, Cabeza Barrera J, Huizinga TW, et al. FcGR genetic polymorphisms and the response to adalimumab in patients with rheumatoid arthritis. Pharmacogenomics. 2015.16(4):373–81.

84. Canhão H, Rodrigues AM, Santos MJ, Carmona-Fernandes D, Bettencourt BF, Cui J, et al. TRAF1/C5 but not PTPRC variants are potential predictors of rheumatoid arthritis response to anti-tumor necrosis factor therapy. Biomed Res Int. 2015.2015:490295.

85. Bek S, Bojesen AB, Nielsen J V, Sode J, Bank S, Vogel U, et al. Systematic review and meta-analysis: pharmacogenetics of anti-TNF treatment response in rheumatoid arthritis. Pharmacogenomics J. 2017.17(5):403–11.

86. Cuppen BVJ, Welsing PMJ, Sprengers JJ, Bijlsma JWJ, Marijnissen ACA, van Laar JM, et al. Personalized biological treatment for rheumatoid arthritis: a systematic review with a focus on clinical applicability. Rheumatology. 2016.55(5):826–39.

87. Gavan S, Harrison M, Iglesias C, Barton A, Manca A, Payne K. Economics of Stratified Medicine in Rheumatoid Arthritis. Curr Rheumatol Rep. 2014.16(12):468.

88. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang J-F, et al. Data mining in healthcare and biomedicine: A survey of the literature. J Med Syst. 2012.36(4):2431–48.

89. Jekic B, Lukovic L, Bunjevacki V, Milic V, Novakovic I, Damnjanovic T, et al. Association of the TYMS 3G/3G genotype with poor response and GGH 354GG genotype with the bone marrow toxicity of the methotrexate in RA patients. Eur J Clin Pharmacol. 2013.69(3):377–83.

90. Wijnen PA, Cremers JP, Nelemans PJ, Erckens RJ, Hoitsma E, Jansen TL, et al. Association of the TNF-alpha G-308A polymorphism with TNF-inhibitor response in sarcoidosis. Eur Respir J. 2014.43(6):1730–9.

91. Iannaccone CK, Lee YC, Cui J, Frits ML, Glass RJ, Plenge RM, et al. Using genetic and clinical data to understand response to disease-modifying anti-rheumatic drug therapy: data from the Brigham and Women's Hospital Rheumatoid Arthritis Sequential Study. Rheumatology (Oxford). 2011.50(1):40–6.

92. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature. 2014.506(7488):376+.

93. Hernandez I, Zhang Y, I. H, Y. Z. Using predictive analytics and big data to optimize pharmaceutical outcomes. Am J Heal Pharm. 2017.74(18):1494–500.

94. Shah SC, Kusiak A. Data mining and genetic algorithm based gene/SNP selection. Artif Intell Med. 2004.31(3):183–96.

95. Zayed N, Awad AB, El-Akel W, Doss W, Awad T, Radwan A, et al. The assessment of data mining for the prediction of therapeutic outcome in 3719 Egyptian patients with chronic hepatitis C. Clin Res Hepatol Gastroenterol. 2013.37(3):254–61.

96. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform. 2002.35(5–6):352–

9.

97. Wessels JAM, van der Kooij SM, le Cessie S, Kievit W, Barerra P, Allaart CF, et al. A clinical pharmacogenetic model to predict the efficacy of methotrexate monotherapy in recent-onset rheumatoid arthritis. Arthritis Rheum. 2007.56(6):1765–75.

98. van Asten F, Rovers MM, Lechanteur YTE, Smailhodzic D, Muether PS, Chen J, et al. Predicting non-response to ranibizumab in patients with neovascular age-related macular degeneration. Ophthalmic Epidemiol. 2014.21(6):347–55.

99. Yang Y, Niehaus KE, Walker TM, Iqbal Z, Walker AS, Wilson DJ, et al. Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. BIOINFORMATICS. 2018.34(10):1666–71.

100. Yin J-Y, Li X-P, Li X-P, Xiao L, Zheng W, Chen J, et al. Prediction models for platinum-based chemotherapy response and toxicity in advanced NSCLC patients. Cancer Lett. 2016.377(1):65–73.

101. Gonzalez Bosquet J, Newtson AM, Chung RK, Thiel KW, Ginader T, Goodheart MJ, et al. Prediction of chemo-response in serous ovarian cancer. Mol Cancer. 2016.15(1):66.

102. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Vol. 112. Springer; 2013.

103. Rashkin SR, Chua KC, Ho C, Mulkey F, Jiang C, Mushiroda T, et al. A Pharmacogenetic Prediction Model of Progression-Free Survival in Breast Cancer using Genome-Wide Genotyping Data from CALGB 40502 (Alliance). Clin Pharmacol Ther. 2019.105(3):738–45.

104. Naushad SM, Dorababu P, Rupasree Y, Pavani A, Raghunadharao D, Hussain T, et al. Classification and regression tree-based prediction of 6-mercaptopurine-induced leucopenia grades in children with acute lymphoblastic leukemia. Cancer Chemother Pharmacol. 2019.83(5):875–80.

105. de Rotte MCFJ, Pluijm SMF, de Jong PHP, Bulatović Ćalasan M, Wulffraat NM, Weel AEAM, et al. Development and validation of a prognostic multivariable model

to predict insufficient clinical response to methotrexate in rheumatoid arthritis. Abu-Shakra M, editor. PLoS One. 2018.13(12):e0208534.

106. Maciukiewicz M, Marshe VS, Hauschild A-C, Foster JA, Rotzinger S, Kennedy JL, et al. GWAS-based machine learning approach to predict duloxetine response in major depressive disorder. J Psychiatr Res. 2018.99:62–8.

107. Jiménez-Sousa MA, Berenguer J, Rallón N, Guzmán-Fulgencio M, López JC, Soriano V, et al. IL28RA polymorphism is associated with early hepatitis C virus (HCV) treatment failure in human immunodeficiency virus-/HCV-coinfected patients. J Viral Hepat. 2013.20(5):358–66.

108. Luxburg U von, Schölkopf B. Statistical Learning Theory: Models, Concepts, and Results. In 2011. p. 651–706.

109. Kureshi N, Abidi SSR, Blouin C. A Predictive Model for Personalized Therapeutic Interventions in Non-small Cell Lung Cancer. IEEE J Biomed Heal informatics. 2016.20(1):424–31.

110. KayvanJoo AH, Ebrahimi M, Haqshenas G. Prediction of hepatitis C virus interferon/ribavirin therapy outcome based on viral nucleotide attributes using machine learning algorithms. BMC Res Notes. 2014.7:565.

111. Bannerman-Thompson H, Bhaskara Rao M, Kasala S. Bagging, Boosting, and Random Forests Using R. In 2013. p. 101–49.

112. Khoshgoftaar TM, Van Hulse J, Napolitano A. Comparing Boosting and Bagging Techniques With Noisy and Imbalanced Data. IEEE Trans Syst Man, Cybern - Part A Syst Humans. 2011.41(3):552–68.

113. Freund Y, Schapire RE, others. Experiments with a new boosting algorithm. In: icml. 1996. p. 148–56.

114. Breiman L. Bagging predictors. Mach Learn. 1996.24(2):123–40.

115. Breiman L. Random Forests. Mach Learn. 2001.45(1):5–32.

116. Genuer R, Poggi J-M, Tuleau-Malot C. Variable selection using random forests. Pattern Recognit Lett. 2010.31(14):2225–36.

117. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intell Syst their Appl. 1998.13(4):18–28.

118. Zhang Y, Wang H, Mao X, Guo Q, Li W, Wang X, et al. A Novel Circulating miRNA-Based Model Predicts the Response to Tripterysium Glycosides Tablets: Moving Toward Model-Based Precision Medicine in Rheumatoid Arthritis. Front Pharmacol. 2018.9(MAY):378.

119. Daemen A, Gevaert O, Ojeda F, Debucquoy A, Suykens JA, Sempoux C, et al. A kernel-based integration of genome-wide data for clinical decision support. Genome Med. 2009.1(4):39.

120. Kim J-W, Sharma V, Ryan ND. Predicting Methylphenidate Response in ADHD Using Machine Learning Approaches. Int J Neuropsychopharmacol. 2015.18(11):pyv052.

121. Vapnik V. The nature of statistical learning theory. Springer science & business media; 2013.

122. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995.20(3):273–97.

123. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys. 1943.5(4):115–33.

124. Caocci G, Baccoli R, Vacca A, Mastronuzzi A, Bertaina A, Piras E, et al. Comparison between an artificial neural network and logistic regression in predicting acute graft-vs-host disease after unrelated donor hematopoietic stem cell transplantation in thalassemia patients. Exp Hematol. 2010.38(5):426–33.

125. Lin E, Hwang Y, Wang S-C, Gu ZJ, Chen EY. An artificial neural network approach to the drug efficacy of interferon treatments. Pharmacogenomics. 2006.7(7):1017–24.

126. Lin C-C, Wang Y-C, Chen J-Y, Liou Y-J, Bai Y-M, Lai I-C, et al. Artificial neural network prediction of clozapine response with combined pharmacogenetic and clinical data. Comput Methods Programs Biomed. 2008.91(2):91–9.

127. Hirose A. Complex-valued neural networks: Advances and applications. Vol. 18. John Wiley & Sons; 2013.

128. Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, et al. Recent advances in convolutional neural networks. Pattern Recognit. 2018.77:354–77.

129. Suzuki K. Artificial neural networks: methodological advances and biomedical applications. BoD--Books on Demand; 2011.

130. Kumar V, Mishra BK, Mazzara M, Thanh DNH, Verma A. Prediction of Malignant and Benign Breast Cancer: A Data Mining Approach in Healthcare Applications. In 2020. p. 435–42.

131. Krishna CL, Reddy PVS. An Efficient Deep Neural Network Multilayer Perceptron Based Classifier in Healthcare System. In: 2019 3rd International Conference on Computing and Communications Technologies (ICCCT). IEEE; 2019. p. 1–6.

132. Kutlay MA, Gagula-Palalic S. Application Of Machine Learning In Healthcare: Analysis On MHEALTH Dataset. Southeast Eur J Soft Comput. 2016.4(2).

133. Naraei P, Abhari A, Sadeghian A. Application of multilayer perceptron neural networks and support vector machines in classification of healthcare data. In: 2016 Future Technologies Conference (FTC). IEEE; 2016. p. 848–52.

134. Sordo M. Introduction to neural networks in healthcare. Open Clin Knowl Manag Med care. 2002.

135. Parr T, Howard J. The matrix calculus you need for deep learning. arXiv Prepr arXiv180201528. 2018.

136. Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. Neurocomputing. 2018.300:70–9.

137. Zhao Z, Morstatter F, Sharma S, Alelyani S, Anand A, Liu H. Advancing feature selection research. ASU Featur Sel Repos Arizona State Univ. 2010.:1–28.

138. Takahashi H, Kaniwa N, Saito Y, Sai K, Hamaguchi T, Shirao K, et al. Construction

of possible integrated predictive index based on EGFR and ANXA3 polymorphisms for chemotherapy response in fluoropyrimidine-treated Japanese gastric cancer patients using a bioinformatic method. BMC Cancer. 2015.15.

139. Chen T, Chen L. Prediction of Clinical Outcome for All Stages and Multiple Cell Types of Non-small Cell Lung Cancer in Five Countries Using Lung Cancer Prognostic Index. EBioMedicine. 2014.1(2–3):156–66.

140. Hanchuan Peng, Fuhui Long, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 2005.27(8):1226–38.

141. Cover TM. The Best Two Independent Measurements Are Not the Two Best. IEEE Trans Syst Man Cybern. 1974.SMC-4(1):116–7.

142. Jain AK, Duin PW, Jianchang Mao. Statistical pattern recognition: a review. IEEE Trans Pattern Anal Mach Intell. 2000.22(1):4–37.

143. DING C, PENG H. MINIMUM REDUNDANCY FEATURE SELECTION FROM MICROARRAY GENE EXPRESSION DATA. J Bioinform Comput Biol. 2005.03(02):185–205.

144. Beh EJ. Simple Correspondence Analysis: A Bibliographic Review. Int Stat Rev. 2007.72(2):257–84.

145. Greenacre M. Correspondence analysis in practice. CRC press; 2017.

146. Greenacre M, Hastie T. The Geometric Interpretation of Correspondence Analysis. J Am Stat Assoc. 1987.82(398):437–47.

147. Kuhn M. Package ‘caret’ [Internet]. CRAN Repository. 2020; p. 1–223. Available from: https://cran.r-project.org/web/packages/caret/caret.pdf

148. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci. 2002.99(10):6562–6.

149. The Cochrane Collaboration. Cochrane Handbook for Systematic Reviews of Interventions [Internet]. Higgins JP, Green S, editors. Chichester, UK: John Wiley & Sons, Ltd; 2008.

150. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009.6(7):e1000097.

151. Kitchenham B, Pearl Brereton O, Budgen D, Turner M, Bailey J, Linkman S. Systematic literature reviews in software engineering – A systematic literature review. Inf Softw Technol. 2009.51(1):7–15.

152. Mariano DCB, Leite C, Santos LHS, Rocha REO, Cardoso De R, Minardi M, et al. A guide to performing systematic literature reviews in bioinformatics. 2017.

153. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med. 2015.162(1):W1.

154. Lin E, Kuo P-H, Liu Y-L, Yu YW-Y, Yang AC, Tsai S-J. A Deep Learning Approach for Predicting Antidepressant Response in Major Depression Using Clinical and Genetic Biomarkers. Front PSYCHIATRY. 2018.9.

155. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet. 2015.16(2):85–97.

156. Shahid M, Choi TG, Nguyen MN, Matondo A, Jo YH, Yoo JY, et al. An 8-gene signature for prediction of prognosis and chemoresponse in non-small cell lung cancer. Oncotarget. 2016.7(52):86561–72.

157. Walter RB, Othus M, Burnett AK, Lowenberg B, Kantarjian HM, Ossenkoppele GJ, et al. Resistance prediction in AML: analysis of 4601 patients from MRC/NCRI, HOVON/SAKK, SWOG and MD Anderson Cancer Center. Leukemia. 2015.29(2):312–20.

158. S.P. L, T.M. B, C.D. H, G. S, C. D. A multimarker model to predict outcome in tamoxifen-treated breast cancer patients. Clin Cancer Res. 2006.12(4):1175–83.

159. Beerenwinkel N, Montazeri H, Schuhmacher H, Knupfer P, von Wyl V, Furrer H, et al. The individualized genetic barrier predicts treatment response in a large cohort of HIV-1 infected patients. PLoS Comput Biol. 2013.9(8):e1003203.

160. Kuo H-C, Wong HS-C, Chang W-P, Chen B-K, Wu M-S, Yang KD, et al. Prediction for Intravenous Immunoglobulin Resistance by Using Weighted Genetic Risk Score Identified From Genome-Wide Association Study in Kawasaki Disease. Circ Cardiovasc Genet. 2017.10(5).

161. Xu M, Tantisira KG, Wu A, Litonjua AA, Chu J, Himes BE, et al. Genome Wide Association Study to predict severe asthma exacerbations in children using random forests classifiers. BMC Med Genet. 2011.12:90.

162. Xie S, Ma W, Shen M, Guo Q, Wang E, Huang C, et al. Clinical and pharmacogenetics associated with recovery time from general anesthesia. Pharmacogenomics. 2018.19(14):1111–23.

163. O'Brien TR, Everhart JE, Morgan TR, Lok AS, Chung RT, Shao Y, et al. An IL28B genotype-based clinical prediction model for treatment of chronic hepatitis C. PLoS One. 2011.6(7):e20904.

164. Khalid S, Khalil T, Nasreen S. A survey of feature selection and feature extraction techniques in machine learning. In: 2014 Science and Information Conference. IEEE; 2014. p. 372–8.

165. Liu H, Motoda H, editors. Feature Extraction, Construction and Selection [Internet]. Boston, MA: Springer US; 1998.

166. Holmes JH. Applying a Learning Classifier System to Mining Explanatory and Predictive Models from a Large Clinical Database. In 2001. p. 103–13.

167. Lever J, Krzywinski M, Altman N. Model selection and overfitting. Nat Methods. 2016.13(9):703–4.

168. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. BMC Med Inform Decis Mak. 2012.12:8.

169. Lin L, Chu H. Quantifying publication bias in meta-analysis. Biometrics. 2018.74(3):785–94.

170. Schrider DR, Kern AD. Supervised Machine Learning for Population Genetics: A New Paradigm. Trends Genet. 2018.34(4):301–12.

171. Bridges M, Heron EA, O'Dushlaine C, Segurado R, Morris D, Corvin A, et al. Genetic Classification of Populations Using Supervised Learning. Kliebenstein DJ, editor. PLoS One. 2011.6(5):e14802.

172. Concato J. Observational versus experimental studies: what's the evidence for a hierarchy? NeuroRx. 2004.1(3):341–7.

173. Boyko EJ. Observational research opportunities and limitations. J Diabetes Complications. 27(6):642–8.

174. Rodin AS, Gogoshin G, Boerwinkle E. Systems biology data analysis methodology in pharmacogenomics. Pharmacogenomics. 2011.12(9):1349–60.

175. Ellis JA, Ong B. The MassARRAY® System for Targeted SNP Genotyping. In 2017. p. 77–94.

176. Mancini MC, Cardoso-Silva CB, Costa EA, Marconi TG, Garcia AAF, De Souza AP. New Developments in Sugarcane Genetics and Genomics. In: Advances of Basic Science for Second Generation Bioethanol from Sugarcane. Cham: Springer International Publishing; 2017. p. 159–74.

177. Fardo DW, Ionita-Laza I, Lange C. On Quality Control Measures in Genome-Wide Association Studies: A Test to Assess the Genotyping Quality of Individual Probands in Family-Based Association Studies and an Application to the HapMap Data. Dermitzakis ET, editor. PLoS Genet. 2009.5(7):e1000572.

178. Fairley S, Lowy-Gallego E, Perry E, Flicek P. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. Nucleic Acids Res. 2020.48(D1):D941–7.

179. Bedoya G, Montoya P, Garcia J, Soto I, Bourgeois S, Carvajal L, et al. Admixture dynamics in Hispanics: A shift in the nuclear genetic ancestry of a South American population isolate. Proc Natl Acad Sci. 2006.103(19):7234–9.

180. Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet. 2006.7(10):781–91.

181. Li W. Three lectures on case control genetic association analysis. Brief Bioinform. 2007.9(1):1–13.

182. Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT. Basic statistical analysis in genetic case-control studies. Nat Protoc. 2011.6(2):121–33.

183. Tang F, Ishwaran H. Random Forest Missing Data Algorithms. Stat Anal Data Min. 2017.10(6):363–77.

184. Rodriguez S, Gaunt TR, Day INM. Hardy-Weinberg Equilibrium Testing of Biological Ascertainment for Mendelian Randomization Studies. Am J Epidemiol. 2009.169(4):505–14.

185. Horita N, Kaneko T. Genetic model selection for a case-control study and a meta-analysis. Meta gene. 2015.5:1–8.

186. Thakkinstian A, McElduff P, D'Este C, Duffy D, Attia J. A method for meta-analysis of molecular association studies. Stat Med. 2005.24(9):1291–306.

187. TYMS thymidylate synthetase [Homo sapiens (human)] [Internet]. NCBI Gene. 2020; Available from: https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=ShowDetailView&TermToSearch=7298

188. ABCC2 ATP binding cassette subfamily C member 2 [Homo sapiens (human)] [Internet]. NCBI Gene. 2020; Available from: https://www.ncbi.nlm.nih.gov/gene/1244

189. Ranganathan P, Culverhouse R, Marsh S, Mody A, Scott-Horton TJ, Brasington R, et al. Methotrexate (MTX) pathway gene polymorphisms and their effects on MTX toxicity in Caucasian and African American patients with rheumatoid arthritis. J Rheumatol. 2008.35(4):572–9.

190. ATIC 5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase [Homo sapiens (human)] [Internet]. NCBI Gene. 2020; Available

from: https://www.ncbi.nlm.nih.gov/gene/471

191. Lee YC, Cui J, Costenbader KH, Shadick NA, Weinblatt ME, Karlson EW. Investigation of candidate polymorphisms and disease activity in rheumatoid arthritis patients on methotrexate. Rheumatology (Oxford). 2009.48(6):613–7.

192. Iannaccone CK, Lee YC, Cui J, Frits ML, Glass RJ, Plenge RM, et al. Using genetic and clinical data to understand response to disease-modifying anti-rheumatic drug therapy: data from the Brigham and Women's Hospital Rheumatoid Arthritis Sequential Study. Rheumatology (Oxford). 2011.50(1):40–6.

193. SLC19A1 solute carrier family 19 member 1 [Homo sapiens (human)] [Internet]. NCBI Gene. 2020; Available from: https://www.ncbi.nlm.nih.gov/gene/6573

194. Owen SA, Hider SL, Martin P, Bruce IN, Barton A, Thomson W. Genetic polymorphisms in key methotrexate pathway genes are associated with response to treatment in rheumatoid arthritis patients. Pharmacogenomics J. 2013.13(3):227–34.

195. MTHFR methylenetetrahydrofolate reductase [Homo sapiens (human)] [Internet]. NCBI Gene. 2020; Available from: https://www.ncbi.nlm.nih.gov/gene/4524

196. Wessels JAM, de Vries-Bouwstra JK, Heijmans BT, Slagboom PE, Goekoop-Ruiterman YPM, Allaart CF, et al. Efficacy and toxicity of methotrexate in early rheumatoid arthritis are associated with single-nucleotide polymorphisms in genes coding for folate pathway enzymes. Arthritis Rheum. 2006.54(4):1087–95.

197. Plaza-Plaza JC, Aguilera M, Cañadas-Garre M, Chemello C, González-Utrilla A, Faus Dader MJ, et al. Pharmacogenetic polymorphisms contributing to toxicity induced by methotrexate in the southern Spanish population with rheumatoid arthritis. OMICS. 2012.16(11):589–95.

198. STEINBROCKER O. THERAPEUTIC CRITERIA IN RHEUMATOID ARTHRITIS. JAMA J Am Med Assoc. 1949.140(8):659.

199. Julià M, Guilabert A, Lozano F, Suarez-Casasús B, Moreno N, Carrascosa JM, et

al. The role of Fcγ receptor polymorphisms in the response to anti–tumor necrosis factor therapy in psoriasis A pharmacogenetic study. JAMA dermatology. 2013.149(9):1033–9.

200. Arandjelovic O. Prediction of health outcomes using big (health) data. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; 2015. p. 2543–6.

201. Ravizza S, Huschto T, Adamov A, Böhm L, Büsser A, Flöther FF, et al. Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. Nat Med. 2019.25(1):57–9.

202. Lee J, An JY, Choi MG, Park SH, Kim ST, Lee JH, et al. Deep Learning-Based Survival Analysis Identified Associations Between Molecular Subtype and Optimal Adjuvant Treatment of Patients With Gastric Cancer. JCO Clin cancer informatics. 2018.2(2):1–14.

203. Liang Z, Huang JX, Zeng X, Zhang G. DL-ADR: a novel deep learning model for classifying genomic variants into adverse drug reactions. BMC Med Genomics. 2016.9(2).

204. Rizk HH, Hamdy NM, Al-Ansari NL, El-Mesallamy HO. Pretreatment Predictors of Response to PegIFN-RBV Therapy in Egyptian Patients with HCV Genotype 4. PLoS One. 2016.11(4):e0153895.

205. Kautzky A, Baldinger P, Souery D, Montgomery S, Mendlewicz J, Zohar J, et al. The combined effect of genetic polymorphisms and clinical parameters on treatment outcome in treatment-resistant depression. Eur Neuropsychopharmacol. 2015.25(4):441–53.

206. Kim S-K, Kim S-Y, Kim J-H, Roh SA, Cho D-H, Kim YS, et al. A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients. Mol Oncol. 2014.8(8):1653–66.

207. Zhang W, Ota T, Shridhar V, Chien J, Wu B, Kuang R. Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. PLoS Comput Biol. 2013.9(3):e1002975.

208. Vidal-Castineira JR, Lopez-Vazquez A, Alonso-Arias R, Moro-Garcia MA, Martinez-Camblor P, Melon S, et al. A predictive model of treatment outcome in patients with chronic HCV infection using IL28B and PD-1 genotyping. J Hepatol. 2012.56(6):1230–8.

209. Neukam K, Camacho A, Caruz A, Rallón N, Torres-Cornejo A, Rockstroh JK, et al. Prediction of response to pegylated interferon plus ribavirin in HIV/hepatitis C virus (HCV)-coinfected patients using HCV genotype, IL28B variations, and HCV-RNA load. J Hepatol. 2012.56(4):788–94.

210. Alterovitz G, Tuthill C, Rios I, Modelska K, Sonis S. Personalized medicine for mucositis: Bayesian networks identify unique gene clusters which predict the response to gamma-D-glutamyl-L-tryptophan (SCV-07) for the attenuation of chemoradiation-induced oral mucositis. Oral Oncol. 2011.47(10):951–5.

211. Kurosaki M, Tanaka Y, Nishida N, Sakamoto N, Enomoto N, Honda M, et al. Pre-treatment prediction of response to pegylated-interferon plus ribavirin for chronic hepatitis C using genetic polymorphism in IL28B and viral factors. J Hepatol. 2011.54(3):439–48.

212. Petrovski S, Szoeke CE, Sheffield LJ, D'souza W, Huggins RM, O'brien TJ. Multi-SNP pharmacogenomic classifier is superior to single-SNP models for predicting drug outcome in complex diseases. Pharmacogenet Genomics. 2009.19(2):147–52.

213. Wu X, Lu C, Ye Y, Chang J, Yang H, Lin J, et al. Germline genetic variations in drug action pathways predict clinical outcomes in advanced lung cancer treated with platinum-based chemotherapy. Pharmacogenet Genomics. 2008.18(11):955–65.

214. Modlich O, Prisack H-B, Munnes M, Audretsch W, Bojar H. Predictors of primary breast cancers responsiveness to preoperative epirubicin/cyclophosphamide-based chemotherapy: translation of microarray data into clinically useful predictive signatures. J Transl Med. 2005.3:32.