TESIS DE LA UNIVERSIDAD DE ZARAGOZA

2024 333

Rocio Aznar Gimeno

Estimación de modelos de clasificación con criterios de optimalidad derivados de la curva ROC

Director/es
Sanz Sáiz, Gerardo
Del Hoyo Alonso, Rafael
Esteban Escaño, Luis Mariano

ISSN 2254-7606

Prensas de la Universidad Universidad Zaragoza

ISSN 2254-7606

Universidad Zaragoza
1542

Tesis Doctoral

ESTIMACIÓN DE MODELOS DE CLASIFICACIÓN CON CRITERIOS DE OPTIMALIDAD DERIVADOS DE LA CURVA ROC

Autor

Rocio Aznar Gimeno

Director/es

Sanz Sáiz, Gerardo
Del Hoyo Alonso, Rafael
Esteban Escaño, Luis Mariano

UNIVERSIDAD DE ZARAGOZA
Escuela de Doctorado

Programa de Doctorado en Matemáticas y Estadística

2024

Universidad Zaragoza
1542

Tesis Doctoral

Título de la tesis

Estimación de modelos de clasificación con criterios de optimalidad derivados de la curva ROC

Autora

Rocío Aznar Gimeno

Directores

Dr. Luis Mariano Esteban Escaño

Dr. Rafael del Hoyo Alonso

Dr. Gerardo Sanz Sáiz

Programa de Doctorado en Matemáticas y Estadística

Universidad de Zaragoza
2024

AGRADECIMIENTOS

Agradezco sinceramente a todas las personas e instituciones que han colaborado, facilitado y brindado apoyo durante todos estos años de doctorado. Tanto desde el punto de vista técnico como personal, su acompañamiento ha sido de gran valor en cada paso de este camino.

En primer lugar, quiero expresar mi más sincero agradecimiento a mis directores de tesis, el Dr. Luis Mariano Esteban, el Dr. Rafael del Hoyo y el Dr. Gerardo Sanz, por motivarme a embarcarme en esta etapa que tanto he disfrutado. Gracias por confiar en mí, incluso más de lo que yo confiaba en mí misma. Gracias también por vuestra orientación, apoyo, acompañamiento y dedicación a lo largo de todo este proceso. Vuestra experiencia y conocimientos han sido fundamentales para mi crecimiento como investigadora y han contribuido significativamente en los logros conseguidos.

Agradecer también al Instituto Tecnológico de Aragón por brindarme facilidades durante este tiempo, permitiéndome conciliar en ocasiones mis responsabilidades laborales con el proceso de doctorado. De igual manera al departamento de Métodos Estadísticos de la Universidad de Zaragoza. El apoyo económico de ambos ha permitido la publicación de las investigaciones y participación en congresos y conferencias, contribuyendo así a la difusión del conocimiento y a mi formación en el área de interés de la tesis.

Agradecer al Instituto de Investigación Sanitaria de Aragón y personal clínico por su compromiso en la mejora de salud de la sociedad. Colaborar con ellos en la resolución de problemas clínicos reales a través de mis investigaciones y trabajos científicos ha sido muy gratificante.

A mis amigas, amigos y familia, por confiar en mí y apoyarme en todas mis decisiones. Gracias por comprender mi ausencia en ciertos momentos y por vuestro cariño incondicional. Este proceso me ha hecho valorar todavía más lo afortunada que soy de teneros en mi vida.

A ti, Gabriel, por lo mismo. Te agradezco tu compañía, tu apoyo en momentos difíciles y tu generosidad al comprender mi compromiso con la tesis.

Quiero expresar especialmente las gracias a mis padres, Paco y Esther. Gracias por haber hecho lo posible cuando era imposible. Gracias por vuestro sacrificio al priorizar nuestros estudios en momentos en los que ha sido muy complicado. Quién soy hoy en día y cada logro que he alcanzado, os lo debo a vosotros. Nunca podré agradecer lo suficiente todo lo que habéis hecho por mí; estas palabras son solo un pequeño reflejo de mi agradecimiento.

Durante estos años de doctorado, son varias las personas queridas que nos han dejado en el camino. Quiero dedicar este trabajo a mis abuelos, Julio y Angelines y, especialmente, a mi abuela Delfina. Una persona apasionada de las matemáticas que nació en una familia pobre y en una época desafiante. Su deseo hubiese sido seguir el mismo camino que yo, y probablemente lo habría hecho mejor, pero no pudo dadas las circunstancias. Gracias, abuela, por todo. Este trabajo es también el tuyo.

Aunque este camino no ha sido fácil, ha merecido la pena. Contribuir al avance científico para la mejora de la calidad de vida de las personas es realmente satisfactorio. Esto es solo el comienzo, seguiremos investigando en esta dirección que tanto he disfrutado.

Compendio de publicaciones

Esta tesis se presenta como compendio de publicaciones con la unidad temática "Estimación de modelos de clasificación con criterios de optimalidad derivados de la curva ROC". Para ello, se hace constar las referencias de los artículos que componen la tesis, ordenadas según son presentados en la memoria:

1. Aznar-Gimeno, R., Esteban, L. M., del-Hoyo-Alonso, R., Borque-Fernando, Á., & Sanz, G. (2022). A Stepwise Algorithm for Linearly Combining Biomarkers under Youden Index Maximization. Mathematics, 10(8), 1221. doi: https://doi.org/10.3390/math10081221.

2. Aznar-Gimeno, R., Esteban, L. M., Sanz, G., del-Hoyo-Alonso, R., & Savirón-Cornudella, R. (2021). Incorporating a new summary statistic into the min–max approach: a min–max–median, min–max–IQR combination of biomarkers for maximising the youden index. Mathematics, 9(19), 2497. doi: https://doi.org/10.3390/math9192497.

3. Aznar-Gimeno, R., Esteban, L. M., Sanz, G., & del-Hoyo-Alonso, R. (2023). Comparing the Min–Max–Median/IQR Approach with the Min–Max Approach, Logistic Regression and XGBoost, Maximising the Youden Index. Symmetry, 15(3), 756. doi: https://doi.org/10.3390/sym15030756.

4. Aznar-Gimeno, R., Esteban, L. M., Labata-Lezaun, G., del-Hoyo-Alonso, R., Abadia-Gallego, D., Paño-Pardo, J. R., Esquillor-Rodrigo M. J., Lanas, Á., & Serrano, M. T. (2021). A clinical decision web to predict ICU admission or death for patients hospitalised with COVID-19 using machine learning algorithms. International Journal of Environmental Research and Public Health, 18(16), 8677. doi: https://doi.org/10.3390/ijerph18168677.

5. Savirón-Cornudella, R., Esteban, L. M., Aznar-Gimeno, R., Pérez-López, F. R., Ezquerro, M. C., Pérez, P. D., Maza, J. M., Sanz, G., Larraz, B. C., & Tajada-Duaso, M. (2020). A cohort study of fetal growth in twin pregnancies by chorionicity: comparison with European and American standards. European

Journal of Obstetrics & Gynecology and Reproductive Biology, 253, 238-248. doi: https://doi.org/10.1016/j.ejogrb.2020.08.044.

6. Gargallo-Puyuelo, C. J., Aznar-Gimeno, R., Carrera-Lasfuentes, P., Lanas, A., Ferrandez, A., Quintero, E., Carrillo, M., Alonso-Abreu, I., Esteban L. M., de la Vega Rodrigálvarez-Chamarro M., Del Hoyo-Alonso, R., & García-González, M. A. (2022). Predictive Value of Genetic Risk Scores in the Development of Colorectal Adenomas. Digestive Diseases and Sciences, 67(8), 4049-4058. doi: https://doi.org/10.1007/s10620-021-07218-5.

7. Aznar-Gimeno, R., Carrera-Lasfuentes, P., del-Hoyo-Alonso, R., Doblaré, M., & Lanas, Á. (2021). Evidence-based selection on the appropriate FIT cut-off point in CRC screening programs in the COVID pandemic. Frontiers in medicine, 8, 712040. doi: https://doi.org/10.3389/fmed.2021.712040.

La memoria de la tesis se estructura de la siguiente manera. En primer lugar, se incluyen unos capítulos introductorios donde se describen los fundamentos teóricos que sustentan las investigaciones realizadas. Posteriormente, se presentan los trabajos publicados y, finalmente, se expone una discusión global y conclusiones derivadas de estas investigaciones.

LUIS MARIANO ESTEBAN ESCAÑO, Doctor en el programa de Métodos Estadísticos por la Universidad de Zaragoza

ACREDITA:

Que ROCÍO AZNAR GIMENO, ha realizado bajo mi dirección el trabajo que les presenta como memoria para optar al grado de Doctor, con el título: “Estimación de modelos de clasificación con criterios de optimalidad derivados de la curva ROC”. Después de su revisión, considero que reúne los requisitos exigidos por la Programa de Doctorado en Matemáticas y Estadística de la Universidad de Zaragoza para ser considerada como Tesis Doctoral por compendio de publicaciones, para ser defendida en sesión pública ante el tribunal que le sea asignado para juzgarla. Y para que conste, de acuerdo con la legislación vigente y a petición de la interesada, firmo el presente certificado en Zaragoza, a 27 de marzo de 2024.

Fdo. Dr. Luis Mariano Esteban Escaño

Rafael del Hoyo Alonso, Doctor en Ciencias Físicas por la Universidad de Zaragoza,

HACE CONSTAR:
Que ROCÍO AZNAR GIMENO, graduada en Matemáticas, ha realizado bajo mi dirección el trabajo que presenta como memoria para optar al grado de Doctor, con el título: "Estimación de modelos de clasificación con criterios de optimalidad derivados de la curva ROC".

A la vista del trabajo, considero que reúne los requisitos exigidos por la Programa de Doctorado en Matemáticas y Estadística de la Universidad de Zaragoza para ser considerada como Tesis Doctoral por compendio de publicaciones y, consiguientemente, para ser defendida en sesión pública ante el tribunal que le sea asignado para juzgarla.

Y para que conste, de acuerdo con la legislación vigente, firmo este documento.

Dr. Rafael del Hoyo Alonso

Haga clic aquí para escribir texto.

Departamento de
Métodos Estadísticos
Universidad Zaragoza

GERARDO SANZ SÁIZ, Doctor en Ciencias Matemáticas por la Universidad de Zaragoza,

HACE CONSTAR:

Que ROCÍO AZNAR GIMENO, graduada en Matemáticas, ha realizado bajo mi dirección el trabajo que presenta como memoria para optar al grado de Doctor, con el título: “Estimación de modelos de clasificación con criterios de optimalidad derivados de la curva ROC”.

A la vista del trabajo, considero que reúne los requisitos exigidos por la Programa de Doctorado en Matemáticas y Estadística de la Universidad de Zaragoza para ser considerada como Tesis Doctoral por compendio de publicaciones y, consiguientemente, para ser defendida en sesión pública ante el tribunal que le sea asignado para juzgarla.

Y para que conste, de acuerdo con la legislación vigente, firmo este documento.

Índice

Lista de Figuras

Lista de Tablas

Capítulo 1

Curva ROC

1.1. Conceptos previos

En el ámbito de la estadística, los problemas de clasificación son un tipo de problema de aprendizaje supervisado donde se estima la pertenencia de la observación o individuo a una categoría o clase, a partir de la información proporcionada por un conjunto de variables explicativas. Cuando el conjunto de categorías posibles es dicotómico, se define como clasificación binaria.

Definición 1. *Sea $\mathcal{X}$ el espacio de variables explicativas e $\mathcal{Y}$ el conjunto de clases posibles. En clasificación binaria, $\mathcal{Y} = \{0, 1\}$. Se define como clasificador o modelo de clasificación a la función $f : \mathcal{X} \to \mathcal{Y}$ que predice la clase $y_i \in \mathcal{Y}$ de una observación i, dada la información de las variables explicativas $\mathbf{x}_i = (x_{i1}, \ldots, x_{in}) \in \mathcal{X}$.*

Los problemas de clasificación binaria son frecuentes en diversos campos, como la medicina, donde abordan, por ejemplo, la detección o diagnóstico de enfermedades, entre otros propósitos. Existen diversas técnicas para abordar los problemas de clasificación, y la elección de la más adecuada depende de los datos y del problema en cuestión. El desafío radica en encontrar el modelo de clasificación óptimo según el criterio de evaluación específico para ese problema. El trabajo de esta tesis se centra principalmente en problemas de clasificación binaria en el ámbito de la salud y la medicina, teniendo como objetivo principal la estimación de modelos de clasificación bajo criterios de optimalidad derivados de la curva ROC (Receiver Operating Characteristic curve, en inglés).

En el contexto de un problema de clasificación binaria, el modelo $f(\mathbf{x}_i)$ suele arrojar un valor predicho continuo. Para decidir en qué clase es clasificado el individuo i se establece un punto de corte c. Si la predicción del modelo supera el punto de corte establecido ($f(\mathbf{x}_i) \geq c$), el individuo se clasifica en una determinada clase; de lo contrario, se clasifica en la otra clase. En lo que sigue, supondremos que si la predicción

supera el punto de corte, el individuo es clasificado en la clase 1 (clase positiva), y en la clase 0 (clase negativa) en caso contrario. En el contexto del diagnóstico médico, la clase positiva se refiere a la categoría que incluye a los pacientes enfermos (presencia del evento), mientras que la clase negativa incluye a los pacientes sanos (ausencia del evento).

Definición 2. *Sea* $\mathbf{x}_i$ *el vector de variables explicativas del individuo i,* y_i *su clase de pertenencia,* $f(\mathbf{x}_i)$ *el clasificador binario y c el punto de corte establecido. Entonces, se define como:*

- *Verdadero positivo: Individuo i que pertenece a la clase positiva (*$y_i = 1$*) y* $\mathbb{1}_{f(\mathbf{x}_i)\geq c} = 1$*. Se denota por VP (o TP, siglas en inglés) al número de verdaderos positivos.*

- *Falso positivo: Individuo i que pertenece a la clase negativa (*$y_i = 0$*) y* $\mathbb{1}_{f(\mathbf{x}_i)\geq c} = 1$*. Se denota por FP al número de falsos positivos.*

- *Verdadero negativo: Individuo i que pertenece a la clase negativa (*$y_i = 0$*) y* $\mathbb{1}_{f(\mathbf{x}_i)\geq c} = 0$ *Se denota por VN (o TN, siglas en inglés) al número de verdaderos negativos.*

- *Falso negativo: Individuo i que pertenece a la clase positiva (*$y_i = 1$*) y* $\mathbb{1}_{f(\mathbf{x}_i)\geq c} = 0$*. Se denota por FN al número de falsos negativos.*

donde $\mathbb{1}$ *denota la función característica o indicadora.*

Para evaluar la capacidad discriminatoria o rendimiento de un clasificador, se definen un conjunto de métricas de evaluación cuyo valor depende del número de verdaderos positivos (VP), falsos positivos (FP), verdaderos negativos (VN) o falsos negativos (FN), dado un punto de corte establecido.

Definición 3. *Denotando por* $\mathbf{X}$ *las variables explicativas, Y la clase real, c el punto de corte establecido y* $f(\mathbf{X})$ *la salida predicha por el modelo. Entonces, se define como:*

- *Tasa de verdaderos positivos (TPR, siglas en inglés) o sensibilidad (Se):* $TPR(c) = \frac{VP}{VP+FN} = P(f(\mathbf{X}) \geq c|Y = 1)$

- *Tasa de verdaderos negativos (TNR, siglas en inglés) o especificidad (Spe):* $TNR(c) = \frac{VN}{VN+FP} = P(f(\mathbf{X}) < c|Y = 0)$

- *Tasa de falsos positivos (FPR, siglas en inglés) o error tipo I:* $FPR(c) = \frac{FP}{VN+FP} = P(f(\mathbf{X}) \geq c|Y = 0)$*=1-especificidad*

- *Tasa de falsos negativos (FNR, siglas en inglés) o error tipo II:*
  $FNR(c) = \frac{FN}{VP+FN} = P(f(\mathbf{X}) < c|Y = 1)$=*1-sensibilidad*
- *Valor predictivo positivo (VPP, siglas en inglés):*
  $VPP(c) = \frac{VP}{VP+FP} = P(Y = 1|f(\mathbf{X}) \geq c)$
- *Valor predictivo negativo (VPN, siglas en inglés):*
  $VPN(c) = \frac{VN}{VN+FN} = P(Y = 0|f(\mathbf{X}) < c)$
- *Accuracy (proporción de predicciones correctas):* $\frac{VP+VN}{VP+FP+FN+VN}$

Es evidente que el clasificador perfecto es aquel que no arroja ningún falso positivo ni falso negativo o, en otras palabras, cuya tasas de verdaderos positivos y negativos (o sensibilidad y especificidad, respectivamente) son igual a 1. Sin embargo, esta situación no suele darse en la realidad, y lo que se busca es maximizar una métrica de evaluación en función del interés y contexto del problema. El punto de corte es ajustado para optimizar el rendimiento del modelo en términos de la métrica de evaluación escogida.

La curva ROC es una herramienta estadística que representa en un solo gráfico los valores de sensibilidad y especificidad para los diferentes puntos de corte, permitiendo visualizar la variación del rendimiento del modelo en función de estos puntos. El concepto fue desarrollado durante la Segunda Guerra Mundial para el análisis de receptores de radar tras el ataque de Pearl Harbor en 1941. El objetivo fue detectar con precisión los aparatos japoneses enemigos en el campo de batalla a partir de sus señales de radar. Según el umbral o punto de corte establecido, el radar podía identificar una amenaza real (considerada como un verdadero positivo), no detectarla (falso negativo), o percibir un ruido de la señal como una amenaza real (falso positivo). Más tarde, a partir de la década de 1960, la curva ROC se hizo popular en estudios de psicología experimental y psicofísica [37, 10, 109]. Su interés también creció en el campo de la medicina debido al potencial de su análisis en la toma de decisiones médicas [62], especialmente al inicio en el campo de la radiología [41].

Desde entonces, la curva ROC ha sido utilizada en diversas disciplinas, especialmente en el ámbito del diagnóstico clínico, y ha sido ampliamente estudiada en la literatura debido a sus propiedades y utilidad como herramienta para evaluar la capacidad de los modelos de clasificación [129, 81]. En la siguiente sección se presenta su definición formal y los resultados derivados de interés.

1.2. Definición y propiedades

Definición 4. *Dado un clasificador binario, sea* F_{Y_k} *la función de distribución de la variable aleatoria* Y_k *que denota los resultados del clasificador para la clase* $k \in \{0, 1\}$

y c el punto de corte. Entonces, se define la curva ROC como la función:

$$\begin{aligned} ROC(c) &= \{(FPR(c), TPR(c)), c \in \mathbb{R}\} \\ &= \{(1 - especificidad(c), sensibilidad(c)), c \in \mathbb{R}\} \\ &= \{(1 - Spe(c), Se(c)), c \in \mathbb{R}\} \\ &= \{(1 - F_{Y_0}(c), 1 - F_{Y_1}(c)), c \in \mathbb{R}\} \end{aligned} \tag{1.1}$$

La definición de la curva ROC permite derivar una serie de propiedades y resultados de interés. Estas propiedades son la base de la estimación de métodos de clasificación con criterios de optimalidad derivados de la curva ROC.

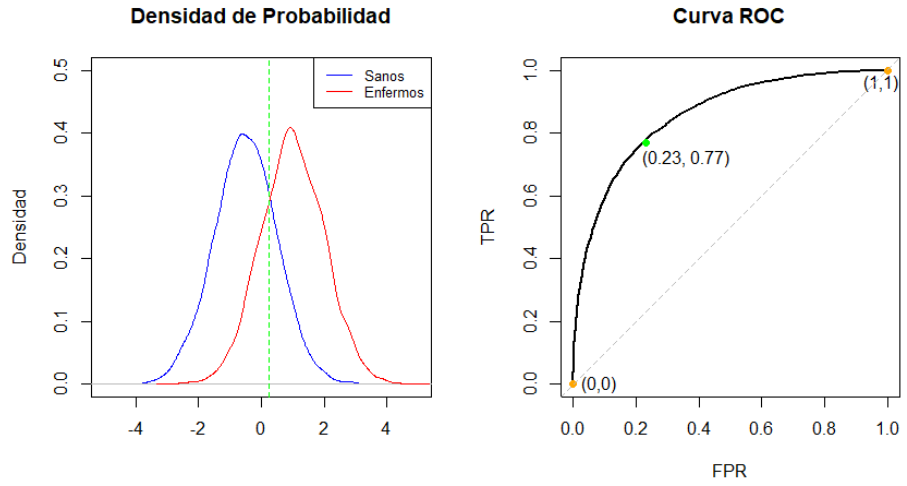
Observación 1. *Los puntos (0,0) y (1,1) son puntos extremos que pertenecen a la curva ROC. El punto (0,0) ocurre para un punto de corte c lo suficientemente grande, de forma que ningún individuo se clasifica como positivo, es decir, no comete errores de falsos positivos (*$\lim_{c\to\infty} FPR(c) = 1 - F_{Y_0}(c) = 0$*) pero tampoco reconoce verderos positivos (*$\lim_{c\to\infty} TPR(c) = 1 - F_{Y_1}(c) = 0$*). Por otro lado, cuando el punto de corte c es lo suficientemente pequeño, el punto (1,1) aparece en la curva ROC, lo que significa que todos los individuos se clasifican como positivos:* $\lim_{c\to-\infty} TPR(c) = 1 - F_{Y_1}(c) = 1 = \lim_{c\to-\infty} FPR(c) = 1 - F_{Y_0}(c) = 1 \implies \lim_{c\to-\infty} ROC(c) = (1, 1)$*. Este resultado representa un clasificador que identifica correctamente todos los verdaderos positivos, pero a costa de aceptar falsos positivos.*

Observación 2. *Cuando la tasa de verdaderos positivos (TPR) es igual a la tasa de falsos positivos (FPR) para todos los posibles valores del punto de corte (*$TPR(c) = FPR(c)\ \forall c$*), la curva ROC es una línea recta que conecta los puntos (0,0) y (1,1), formando la diagonal del espacio* $[0, 1] \times [0, 1]$*. En esta situación, la curva ROC representa un clasificador aleatorio o sin capacidad discriminativa. Por otro lado, si la curva ROC pasa por el punto (0,1), se dice que la clasificación es perfecta, ya que identifica correctamente todos los verdaderos positivos sin cometer errores de falsos positivos.*

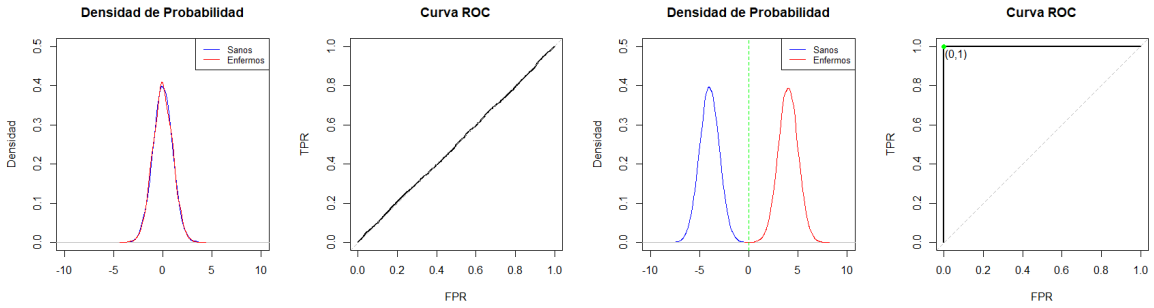
En la Figura 1.1 se presentan ejemplos de curvas ROC correspondientes a conjuntos de datos con diferentes características en las distribuciones de la población sana y enferma. La curva ROC describe la capacidad para separar las distribuciones. En concreto, la Figura 1.1a muestra la curva ROC para un ejemplo de datos, donde las distribuciones de los individuos sanos (curva azul) y los enfermos (curva roja) se superponen en ciertas áreas. En tales casos, independientemente del punto de corte elegido, el clasificador no logrará ser perfecto. Existen diferentes técnicas para seleccionar el punto de corte óptimo, que se discuten en el capítulo 3. Asumiendo que

tiene la misma importancia clasificar correctamente tanto a los enfermos como a los sanos (línea vertical verde), el punto de la curva ROC correspondiente para ese punto de corte se representa en verde.

Las Figuras 1.1b-1.1c ilustran casos opuestos extremos. En la primera, se observa una curva ROC que resulta de distribuciones que se superponen totalmente, indicando la incapacidad del clasificador para discriminar entre las clases. En cambio, en el segundo caso, las distribuciones son completamente distinguibles, lo que se refleja en una curva ROC que representa una clasificación perfecta.



(a) Poblaciones parcialmente solapadas



(b) Poblaciones totalmente solapadas

(c) Poblaciones no solapadas

Figura 1.1: Ejemplos de curvas ROC

Proposición 1. *La curva ROC es una función monótona creciente que toma valores en* $[0,1] \times [0,1]$.

Demostración. La demostración es directa por definición, dada la relación existente entre la tasas de falsos y verdaderos positivos. Tanto la tasa de falsos positivos ($FPR(c) = 1 - F_{Y_0}(c)$) como la tasa de verdaderos positivos ($TPR(c) = 1 - F_{Y_1}(c)$) disminuyen tendiendo a 0 a medida que el punto de corte c aumenta, y ambas crecen acercándose a 1 cuando c tiende a $-\infty$.

□

Proposición 2. *La curva ROC puede ser definida de forma equivalente a (1.1) como:*

$$ROC(t) = 1 - F_{Y_1}(F_{Y_0}^{-1}(1-t)),\ t \in [0,1] \tag{1.2}$$

$$= \overline{F}_{Y_1}(\overline{F}_{Y_0}^{-1}(t)),\ t \in [0,1] \tag{1.3}$$

donde $F_{Y_0}^{-1}(t) = \inf\{x \in \mathbb{R} : F_{Y_0}(x) \geq t\}$ *es la función de distribución inversa y* $\overline{F}_{Y_k} = 1 - F_{Y_k}$ *la función de supervivencia de* $F_{Y_k}, k = 0, 1$.

Demostración. Utilizando la forma explícita de la curva, considerando $t = 1 - F_{Y_0}(c)$ y sustituyendo en (1.1), se tiene (1.2). Aplicando la definición de $F_{Y_0}^{-1}$ sobre (1.2), se obtiene (1.3). □

Observación 3. *La monotonía creciente de la curva ROC, como se establece en la Proposición 1, también se evidencia directamente a partir de su definición equivalente (1.3). Dado que la función de supervivencia es decreciente, su inversa también lo es, y al componerlas se obtiene una función monótona creciente: sea* $t_1 < t_2$*, entonces* $\overline{F}_{Y_1}(t_2) \leq \overline{F}_{Y_1}(t_1)$ *y* $\overline{F}_{Y_0}^{-1}(t_2) \leq \overline{F}_{Y_0}^{-1}(t_1) \implies \overline{F}_{Y_1}(\overline{F}_{Y_0}^{-1}(t_1)) \leq \overline{F}_{Y_1}(\overline{F}_{Y_0}^{-1}(t_2))$.

Proposición 3. *La curva ROC es invariante a transformaciones monótonas crecientes.*

Demostración. Sea g una función monótona creciente. La demostración es inmediata ya que la función de distribución de Y_k, $k = 0, 1$, es la misma que la de $g(Y_k)$ con el punto de corte $d = g(c)$: $1 - F_{Y_k}(c) = P(Y_k > c) = P(g(Y_k) > g(c)) = P(g(Y_k) > d) = 1 - F_{g(Y_k)}(d)$. □

De acuerdo con la definición de la curva ROC que aparece en (1.1), es evidente inferir que para representarla se requiere evaluar la tasa de verdaderos positivos y la tasa de falsos positivos para todos los valores posibles del punto de corte $c \in \mathbb{R}$, o bien conocer las distribuciones F_{Y_0} y F_{Y_1}. En consecuencia, se presentan a continuación las estimaciones de la curva ROC, tanto paramétrica como no paramétrica.

1.3. Estimación paramétrica

La estimación paramétrica de la curva ROC asume que la variable respuesta o de decisión Y_k sigue una distribución perteneciente a una familia de distribuciones, que puede variar entre las clases $k = 0, 1$. Este tipo de estimaciones se simplifica en la tarea de estimar los parámetros correspondientes a las distribuciones que se asume

que sigue la variable de respuesta. Por lo tanto, resulta esencial conocer o seleccionar adecuadamente la distribución que sigue dicha variable.

Considerando variables de respuesta continuas, el caso binormal (donde las distribuciones de los resultados del clasificador para la clase positiva y negativa siguen distribuciones normales) es un caso de interés y el más usado en la estimación de la curva ROC [70]. Esto se debe a dos resultados claves. En primer lugar, aunque las variables no sigan una distribución normal, muchas de ellas se pueden ajustar a una distribución normal mediante transformaciones monótonas. Zou et al. [132] proponen transformaciones del tipo Box-Cox [16] para la transformación de los datos a la binormalidad. En segundo lugar, la curva ROC es invariante ante tranformaciones monótonas, lo que significa que la estimación de la curva ROC para una variable respuesta que sigue una distribución normal tras transformación monótona es equivalente a estimar la curva ROC considerando el caso binormal. La asunción de normalidad de las variables transformadas, sin especificar la transformación monótona que lo permite, es un enfoque comúnmente utilizado en lo que se conoce en la literatura como estimación semiparamétrica de la curva ROC [43, 18].

A continuación, presentamos la estimación (semi-)paramétrica de la curva ROC con el enfoque binormal. Supongamos que $Y_0 \sim N(\mu_0, \sigma_0)$ y $Y_1 \sim N(\mu_1, \sigma_1)$ con $\mu_1 > \mu_0$, o equivalentemente, $\frac{Y_k - \mu_k}{\sigma_k} \sim N(0,1), k = 0, 1$. Sea c el punto de corte y F_{Y_k} la función de distribución de Y_k, entonces $1 - F_{Y_k}(c) = 1 - \Phi\left(\frac{c-\mu_k}{\sigma_k}\right) = \Phi\left(\frac{\mu_k - c}{\sigma_k}\right)$, donde Φ representa la función de distribución de la normal estándar. Considerando la forma explícita de la curva ROC (1.2) con $t = 1 - F_{Y_0}(c)$, se obtiene la expresión de la curva ROC para el caso binormal:

$$
\begin{aligned}
ROC(t) &= 1 - F_{Y_1}(F_{Y_0}^{-1}(1-t)) && (1.4)\\
&= \Phi\left(\frac{\mu_1 - F_{Y_0}^{-1}(1-t)}{\sigma_1}\right) && (1.5)\\
&= \Phi\left(\frac{\mu_1 - \mu_0 + \sigma_0 \Phi^{-1}(t)}{\sigma_1}\right) && (1.6)\\
&= \Phi\left(a + b\Phi^{-1}(t)\right), \ t \in [0,1] && (1.7)
\end{aligned}
$$

donde $a = \frac{\mu_1 - \mu_0}{\sigma_1}$ y $b = \frac{\sigma_0}{\sigma_1}$.

La expresión de la curva ROC binormal estimada $\widehat{ROC}(t)$ de (1.7) se puede obtener estimando los parámetros de la distribución mediante el método de máxima verosimilitud: $\hat{a} = \frac{\hat{\mu}_1 - \hat{\mu}_0}{\hat{\sigma}_1}$ y $\hat{b} = \frac{\hat{\sigma}_0}{\hat{\sigma}_1}$.

La naturaleza robusta del estimador binormal de la curva ROC ha sido discutida en la literatura [110, 40, 119, 36]. Algunos autores han respaldado la robustez del estimador binormal, como Hanley [40], que respalda las conclusiones de Swets [110],

argumentando que, incluso con la presencia de algunas observaciones que no siguen una distribución normal, el modelo ofrece buenos resultados. Sin embargo, Walsh [119] cuestiona esta afirmación, discutiendo la capacidad del estimador binormal para producir inferencias válidas en circunstancias en las que los datos no cumplen con la suposición de normalidad. A partir de un estudio de simulación, concluye que dicho estimador resulta sensible a la especificación errónea del modelo. En los últimos años, se ha prestado mucha atención al impacto de los errores en la especificación del modelo en el campo de la salud y resulta muy importante su estimación. Gonçalves et al. [36] proporcionan una revisión de la literatura más detallada sobre la robustez del estimador binormal.

Intervalos de confianza

Una vez se ha realizado la estimación, se pueden generar intervalos de confianza que permitan establecer, con cierto nivel de confianza, el rango en el cual se espera que se encuentre la verdadera curva ROC.

Denotando $d = \Phi^{-1}(t)$, $R_d = \hat{a} + \hat{b}d$ y siendo α el nivel de confianza, para obtener el intervalo de confianza puntual del $100(1-\alpha)\,\%$ de la curva ROC (1.7), se requiere una estimación de la varianza:

$$\widehat{Var}(R_d) = \widehat{Var}(\hat{a}) + d^2\widehat{Var}(\hat{b}) + 2d\widehat{Cov}(\hat{a}, \hat{b}) \tag{1.8}$$

Zhou et al. [129] y Obuchowski et al. [76] presentan las siguientes fórmulas de las estimaciones de la varianza y covarianza de $\hat{a}$ y $\hat{b}$ para el caso binormal (original o tras transformación):

$$\widehat{Var}(\hat{a}) = \frac{n_1(\hat{a}^2 + 2) + 2n_0\hat{b}^2}{2n_0n_1} \tag{1.9}$$

$$\widehat{Var}(\hat{b}) = \frac{(n_1 + n_0)\hat{b}^2}{2n_0n_1} \tag{1.10}$$

$$\widehat{Cov}(\hat{a}, \hat{b}) = \frac{\hat{a}\hat{b}}{2n_0} \tag{1.11}$$

donde n_0 y n_1 denotan el número de individuos de la clase negativa (sanos) y positiva (enfermos), respectivamente.

Eugene Demidenko [26] propone un intervalo confianza puntual para la curva ROC binormal, utilizando los estimadores de momento insesgados para las varianzas: $Var(\bar{x}_k) = \frac{\sigma_k^2}{n_k}$, $Var(s_k^2) = \frac{2\sigma_k^4}{n_k - 1}$, $k = 0, 1$. Dado $X_d = \frac{\bar{x}_1 - \bar{x}_0 + s_0 d}{s_1}$ y siendo $\bar{x}_1 - \bar{x}_0$, s_0^2 y s_1^2 independientes, mediante el método delta [77], obtiene la siguiente expresión de la varianza de X_d:

$$\widehat{Var}(X_d) = \frac{1}{s_1^2}\left(\frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}\right) + \frac{d^2 s_0^2}{2s_1^2(n_0 - 1)} + \frac{(\bar{x}_1 + \bar{x}_0 - s_0 d)^2}{2s_1^2(n_1 - 1)} \tag{1.12}$$

A partir de esta expresión, el autor presenta el intervalo de confianza puntual para la curva ROC binormal, con d fijo: $\Phi(X_d \pm z_{1-\alpha/2}\sqrt{\widehat{Var}(X_d)})$, donde z denota el cuantil de la normal estándar.

1.4. Estimación no paramétrica

La estimación no paramétrica surge de manera natural como una alternativa útil en situaciones donde resulta difícil conocer la distribución subyacente de la variable respuesta de cada grupo, lo que suele ser común en la práctica. Esto se debe a que no requiere hacer suposiciones sobre dicha distribución, lo que la hace más flexible y robusta que la estimación paramétrica [33], siendo además menos sensible a datos atípicos.

A continuación se presentan dos enfoques para la estimación no paramétrica de la curva ROC. En primer lugar, se describe el método empírico, el cual utiliza la función de distribución empírica a partir de los datos. En segundo lugar, se introduce como alternativa el método tipo Kernel, que proporciona una estimación suavizada de la curva ROC.

1.4.1. Método empírico

Sea $c \in \mathbb{R}$ un punto de corte seleccionado, n_k el número de individuos de la clase k e y_{ki} la salida del clasificador para el individuo $i = 1, \ldots, n_k$ de la clase k, entonces las funciones de distribución empíricas se definen como:

$$\hat{F}_{Y_k}(c) = \frac{1}{n_k}\sum_{i=1}^{n_k} I(y_{ki} \leq c), \quad k = 0, 1. \tag{1.13}$$

donde I denota la función indicadora.

Sustituyendo (1.13) en la definición de la curva ROC (1.1) se obtiene la expresión de la curva ROC empírica:

$$\widehat{ROC}(c) = \left\{ \left(\frac{1}{n_0}\sum_{i=1}^{n_0} I(y_{0i} > c), \frac{1}{n_1}\sum_{i=1}^{n_1} I(y_{1i} > c) \right), \ c \in \mathbb{R} \right\} \tag{1.14}$$

La curva ROC empírica se traza en el plano como la unión lineal de cada punto obtenido para cada valor de c, dando como resultado una curva escalonada, a diferencia del trazado suave que ofrece la estimación paramétrica.

La función de distribución empírica $\hat{F}_{Y_k}$ es un estimador consistente [115, 43] que converge a la función de distribución verdadera F_{Y_k} a medida que el tamaño de la muestra aumenta. Por tanto, cuando se utilicen tamaños de muestras grandes, la curva ROC estimada será más suave. Sin embargo, para tamaños de muestra especialmente

pequeños o, equivalentemente, un pequeño número de puntos de corte considerados, la curva estará formada por un número reducido de puntos, lo que resultará en un trazado no suave e irregular.

Intervalos de confianza

Considerando la estimación empírica (1.14), la sensibilidad y especificidad se expresan como una proporción y, por tanto, se pueden aplicar métodos para el calculo de intervalos de confianza de una proporción binomial. El análisis y estimación de intervalos de confianza para proporciones binomiales ha sido una línea de estudio en la literatura, donde diversos autores han propuesto diferentes métodos para el cálculo de estos intervalos de confianza [120, 22, 1, 17, 130, 118]. A continuación se presentan algunos de los más conocidos.

El intervalo estándar es el denominado intervalo de Wald o intervalo asintótico, que se basa en la distribución asintótica del estimador de la proporción muestral $\hat{p}$ y tiene la siguiente forma conocida: $\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, donde n es el tamaño de la muestra y z el cuantil de la distribución normal estándar. A pesar de su simplicidad de cálculo, esta aproximación no funciona bien, especialmente cuando el tamaño de muestra n es pequeño o la probabilidad de éxito p está cercana a 0 o 1 [17]. Una alternativa al intervalo de Wald es el intervalo score o de Wilson [120, 1]. Se construye invirtiendo las aproximaciones del teorema central del límite y resolviendo la ecuación de segundo grado resultante. Su intervalo de confianza para la sensibilidad Se (el de la especificidad es similar) tiene la siguiente expresión: $\frac{\hat{Se}+z^2_{1-\alpha/2}/(2n_1)\pm z_{1-\alpha/2}\sqrt{[\hat{Se}(1-\hat{Se})+z^2_{1-\alpha/2}/(4n_1)]/n_1}}{1+z^2_{1-\alpha/2}/n_1}$. Agresti y Coull [1] propusieron una modificación al intervalo de Wald (intervalo de Wald ajustado) que solventa sus limitaciones. Considerando $\tilde{n} = n + z^2_{1-\alpha/2}$ y $\tilde{p} = \frac{n\hat{p}+\frac{z^2_{1-\alpha/2}}{2}}{\tilde{n}}$ y sustituyendo n por $\tilde{n}$ y $\hat{p}$ por $\tilde{p}$ en la fórmula del intervalo de Wald, se obtiene el intervalo de confianza propuesto. Clopper y Pearson [22] proponen un intervalo de confianza, denominado intervalo de confianza exacto de Clopper-Pearson, que se basa directamente en la distribución binomial. Aunque este intervalo de confianza puede ser útil en situaciones donde el tamaño de muestra sea demasiado pequeño y no puedan aplicarse con confianza los métodos anteriores, es conservador y alcanza un nivel de cobertura igual o superior al nominal, lo que limita su utilidad práctica [1].

Otra alternativa es el enfoque bootstrap [27, 28] para el cálculo de los intervalos de confianza [86, 65]. Los más populares en la práctica son el método bootstrap percentil y el corregido de sesgo acelerado (BCa, siglas en inglés) [129, 46].

1.4.2. Método tipo kernel

El enfoque no paramétrico de tipo núcleo o kernel proporciona una estimación suavizada de la curva ROC. Este enfoque fue propuesto por Zou et al. [131] a través de la estimación de las distribuciones de la variable de decisión a partir de funciones de densidad kernel. A partir de estas funciones estimadas, se obtienen las funciones de distribución y, consecuentemente, la curva ROC estimada.

Definición 5. *Sean $y_{k1}, \ldots, y_{kn_k}$ el conjunto de datos para el grupo $k = 0, 1$. Entonces, se definen las funciones de densidad kernel estimadas de cada grupo como:*

$$\hat{f}_k(x) = \frac{1}{n_k h_k} \sum_{i=1}^{n_k} K_k \left(\frac{x - y_{ki}}{h_k} \right) \tag{1.15}$$

donde h_k denota el conocido como ancho de banda y K_k la función núcleo o kernel. El ancho de banda h determina la cantidad de suavidad de la curva y las funciones kernel cumplen que $\int_{\mathbb{R}} K_k(x)dx = 1$, $\int_{\mathbb{R}} xK_k(x)dx = 0$ y $\int_{\mathbb{R}} x^2 K_k(x)dx > 0$.

A partir de las funciones de densidad estimadas de (1.15), las funciones de distribución pueden estimarse como:

$$\hat{F}_{Y_k}(x) = \frac{1}{n_k} \sum_{i=1}^{n_k} \int_{-\infty}^{x} \frac{1}{h_k} K_k \left(\frac{t - y_{ki}}{h_k} \right) dt, \ k = 0, 1. \tag{1.16}$$

Por tanto, la elección de la función kernel y, especialmente, el ancho de banda determinará la estimación kernel de la curva ROC. Zou et al. [131] sugieren utilizar el kernel biweight que ofrece la ventaja de ser bastante suave en un intervalo finito:

$$K(x) = \frac{15}{16}(1 - x^2)^2, \ x \in (-1, 1) \tag{1.17}$$

Sin embargo, la elección más popular es el kernel gaussiano [132, 101]. Concretamente, los estudios presentados en esta tesis consideran dicho kernel. En este caso, las funciones de distribución estimadas se expresan como:

$$\widehat{F}_{Y_k}(x) = \frac{1}{n_k h_k} \sum_{i=1}^{n_k} \Phi \left(\frac{x - y_{ki}}{h_k} \right), \ k = 0, 1. \tag{1.18}$$

donde Φ es la función de distribución de una normal estándar.

La elección del ancho de banda h adecuado es un aspecto clave del estimador que ha sido ampliamente estudiado en la literatura [128, 39, 45]. Cuanto menor sea el valor de h, el número de observaciones involucradas en la estimación será menor y la función puede estar menos suavizada. En estos casos, el sesgo se reduce pero aumenta la varianza. Se debe encontrar un equilibrio entre el sesgo y varianza con el objetivo de seleccionar el parámetro de suavizado óptimo.

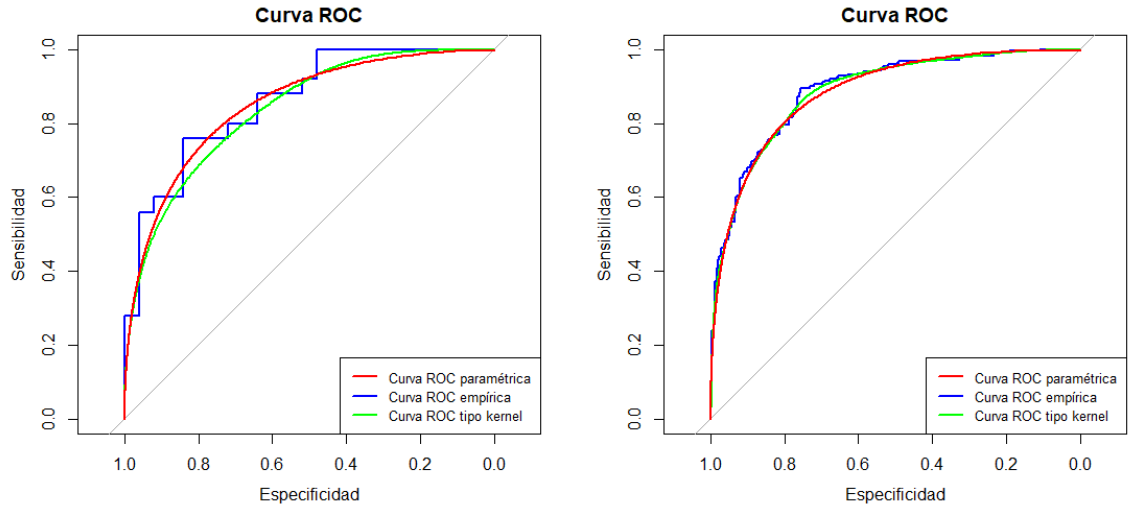
Zou et al. [131] sugieren utilizar el siguiente ancho de banda genérico [90, 79, 104]:

$$h_k = 0{,}9 \,\text{mín}\left(SD_k, IQR_k/1{,}34\right) n_k^{-0{,}2} \tag{1.19}$$

donde SD_k y IQR_k denotan la desviación estándar y el rango intercuartílico, respectivamente, del grupo $k = 0, 1$.

En resumen, el método tipo kernel es una alternativa útil para la estimación de la curva ROC, puesto que permite obtener una curva suave y es menos sensible al ruido que el método empírico. Asimismo, su estimación no depende de supuestos de distribución, como ocurre con la estimación paramétrica.

En la Figura 1.2 se representan las estimaciones paramétricas, empíricas y suavizadas tipo kernel de la curva ROC para un conjunto de datos normales con distintos tamaños de muestra ($n_0 = n_1 = 25$ y $n_0 = n_1 = 250$). Para la estimación de tipo kernel, se ha utilizado el kernel gaussiano y el ancho de banda genérico (1.19). Se observa el trazado suave de las estimaciones paramétricas y de tipo kernel, en contraste con el trazado escalonado de la curva ROC empírica, especialmente para tamaños de muestra más pequeños (Figura 1.2a), donde los escalones son más pronunciados. Para tamaños de muestra más grandes (Figura 1.2b), el trazado de la curva ROC empírica es más suave y las curvas estimadas muestran una mayor similitud.



(a) $n_0 + n_1 = 50$ (b) $n_0 + n_1 = 500$

Figura 1.2: Ejemplos de estimaciones de la curva ROC

Intervalos de confianza

Zou et al. [131] presentan intervalos de confianza para la tasa de verdaderos positivos (TPR) dado un valor de 1-especificidad (o FPR) específico, y una región de confianza para el punto de la curva ROC (FPR, TPR) dado un valor de punto

de corte fijo. Los autores proponen construir intervalos de confianza en un espacio *logit* y luego transformar los resultados de vuelta al espacio ROC, considerando que las aproximaciones normales funcionan mejor en un espacio no restringido con menos sesgo. En concreto, sea $u = logit(FPR) = log(\frac{FPR}{1-FPR})$ y $v = logit(TPR)$ pertenecientes al espacio *logit*. Entonces, $FPR = logit^{-1}(u) = \frac{1}{1+\exp(-u)}$ y $TPR = logit^{-1}(v)$.

Nos centramos en la construcción de la región de confianza para (FPR, TPR) en un umbral fijo. En el espacio *logit* las estimaciones $\hat{u}$ y $\hat{v}$ son asintóticamente normales e independientes con medias u y v, respectivamente, y con varianzas estimadas:

$$\widehat{Var(\hat{u})} = \left(n_0 \widehat{FPR}(1-\widehat{FPR})\right)^{-1} = \left(n_0 \widehat{Spe}(1-\widehat{Spe})\right)^{-1} \tag{1.20}$$

$$\widehat{Var(\hat{v})} = \left(n_1 \widehat{TPR}(1-\widehat{TPR})\right)^{-1} = \left(n_1 \widehat{Se}(1-\widehat{Se})\right)^{-1} \tag{1.21}$$

donde Se y Spe denotan la sensibilidad y especificidad, respectivamente. Por tanto, los intervalos de confianza al $100(1-\alpha)\,\%$ para u y v vienen dados por $\hat{u} \pm z_{\alpha/2}\sqrt{\widehat{Var(\hat{u})}}$ y $\hat{v} \pm z_{\alpha/2}\sqrt{\widehat{Var(\hat{v})}}$, respectivamente. Estos intervalos definen una región de confianza (rectángulo) de (u, v) con un nivel de confianza de $100(1-\alpha)^2\,\%$, por ser $\hat{u}$ y $\hat{v}$ independientes. Finalmente, volviendo al espacio ROC, se obtiene el rectángulo de confianza para el punto de la curva (FPR, TPR) correspondiente a un punto de corte dado, que viene definido por los puntos combinación de los valores:

$$logit^{-1}(\hat{u} \pm z_{\alpha/2}\sqrt{\widehat{Var(\hat{u})}}) \tag{1.22}$$

$$logit^{-1}(\hat{v} \pm z_{\alpha/2}\sqrt{\widehat{Var(\hat{v})}}) \tag{1.23}$$

Hall et al. [38] también construyen intervalos de confianza de la curva ROC utilizando estimaciones kernel. En concreto, sugieren métodos asintóticos y presentan un enfoque alternativo basado en bootstrap. Otros autores, como Bertail et al. [13], también proponen enfoques bootstrap para obtener intervalos de confianza para la curva ROC. En su caso, sugieren un método de remuestreo llamado 'bootstrap suavizado'. Más recientemente, Martinez-Camblor et al. [66] estudiaron métodos no paramétricos para la construcción de bandas de confianza para la curva ROC, presentando un nuevo método que utiliza una técnica de bootstrap y se basa en un enfoque de suavizado para la estimación de la curva ROC.

1.5. Comparación de curvas ROC

El objetivo final de un problema de clasificación es seleccionar el modelo que mejor prediga la clase de pertenencia de la observación basándose en unas características o atributos observables. En otras palabras, seleccionar el modelo que menor error cometa

en la predicción y alcance el mayor rendimiento entre un conjunto de candidatos. Para ello, es necesario disponer de métodos que permitan comparar el rendimiento o capacidad predictiva de dos modelos de clasificación con el fin de seleccionar el óptimo.

La curva ROC es una herramienta que permite evaluar la capacidad de discriminación de los modelos de clasificación que ha sido presentada y formulada en las secciones anteriores. Dadas las estimaciones de la curva ROC presentadas, esta sección aborda la comparación de curvas ROC desde principalmente dos enfoques: comparación de curvas ROC bajo asunción de binormalidad (estimación paramétrica) y comparación de valores de sensibilidad o especificidad para un punto de corte dado (evaluación empírica). La comparación se aborda a través de tests estadísticos con contraste de hipótesis de igualdad.

Comparación de curvas ROC

Aunque en ocasiones la diferencia entre dos curvas ROC puede ser fácilmente observada por representación gráfica, realizar una prueba de hipótesis para aceptar o rechazar la igualdad de las curvas ROC asegura una conclusión más sólida para todas las situaciones.

Suponiendo el caso binormal, la estimación de la curva ROC (1.7) viene determinada por los parámetros a y b. Por tanto, el test de comparación de curvas ROC se resume en:

$$H_0 : a_1 = a_2 \ y \ b_1 = b_2$$
$$H_1 : a_1 \neq a_2 \ o \ b_1 \neq b_2$$

donde a_j es el parámetro a que caracteriza la curva ROC del clasificador $j = 1, 2$. Metz et al. [71, 72] presentan el estadístico que resuelve el contraste:

$$\chi^2 = \frac{\hat{a}_{12} Var(\hat{b}_{12}) + \hat{b}_{12}^2 Var(\hat{a}_{12}) - 2\hat{a}_{12}\hat{b}_{12} Cov(\hat{a}_{12}, \hat{b}_{12})}{Var(\hat{a}_{12}) Var(\hat{b}_{12}) - Cov(\hat{a}_{12}, \hat{b}_{12})^2} \tag{1.24}$$

que sigue una distribución chi-cuadrado de dos grados de libertad, donde $a_{12} = a_1 - a_2$ y $b_{12} = b_1 - b_2$. Las expresiones de las varianzas y covarianzas difieren dependiendo de si las muestras son independientes o pareadas. En general, suele interesar comparar la capacidad predictiva de dos clasificadores bajo el mismo conjunto de pacientes. En este caso, las varianzas y covarianzas vienen dadas por:

$$Var(\hat{a}_{12}) = Var(\hat{a}_1 - \hat{a}_2) = \hat{\sigma}^2_{a_1} + \hat{\sigma}^2_{a_2} - 2\hat{\sigma}_{a_1 a_2} \tag{1.25}$$
$$Var(\hat{b}_{12}) = Var(\hat{b}_1 - \hat{b}_2) = \hat{\sigma}^2_{b_1} + \hat{\sigma}^2_{b_2} - 2\hat{\sigma}_{b_1 b_2} \tag{1.26}$$
$$Cov(\hat{a}_{12}, \hat{b}_{12}) = Cov(\hat{a}_1 - \hat{a}_2, \hat{b}_1 - \hat{b}_2) = \hat{\sigma}_{a_1 b_1} + \hat{\sigma}_{a_2 b_2} - \hat{\sigma}_{a_1 b_2} - \hat{\sigma}_{a_2 b_1} \tag{1.27}$$

donde $\hat{\sigma}^2_{a_j}$ es la varianza estimada de $\hat{a}_j$, $\hat{\sigma}^2_{b_j}$ la varianza estimada de $\hat{b}_j$ y $\hat{\sigma}_{a_j b_j}$ la covarianza estimada de $\hat{a}_j$ y $\hat{b}_j$.

Venkatraman y Begg [113, 112] proponen métodos alternativos para evaluar la igualdad de las curvas ROC, sin necesidad de supuesto de binormalidad. Los métodos se basan en test de permutaciones, que se apoyan en la idea de que si no hay diferencia real, entonces las permutaciones aleatorias deberían producir distribuciones de estadísticas similares.

Comparación de sensibilidad/especificidad

La optimización de la capacidad predictiva de un modelo depende de las condiciones y el contexto específico del problema. Por ejemplo, se puede preferir una alta tasa de verdaderos positivos (sensibilidad) mientras se mantiene un nivel mínimo de tasa de verdaderos negativos (especificidad), o viceversa. Esto es especialmente relevante en situaciones diagnósticas, donde se busca asegurar la detección precisa de pacientes enfermos. En estos casos, se establece un umbral mínimo de sensibilidad y se comparan los clasificadores para seleccionar el que logre la máxima especificidad. En ciertos escenarios, puede existir la situación opuesta donde se prefiere garantizar un mínimo porcentaje de pacientes sanos correctamente clasificados (alta especificidad). El motivo de esto pueden ser las consecuencias graves que conlleva la intervención en pacientes sanos o las limitaciones de recursos para el tratamiento o seguimiento. Una situación reciente que experimentó ambas situaciones fue la pandemia del COVID-19. En las primeras olas se experimentó escasez de recursos, lo que limitó la capacidad de atención y tratamiento de los pacientes. Sin embargo, en olas posteriores, se observó una mayor capacidad de atención, lo que resultó en una mayor probabilidad de ingreso de pacientes, incluso con características similares a los de las primeras olas [9].

En estos casos, se puede realizar una comparación de valores de sensibilidad/especificidad dado un valor específico de especificidad/sensibilidad. Al tratarse de una comparación de proporciones, se pueden utilizar pruebas estadísticas para la comparación de proporciones, concretamente para datos apareados, debido a que el objetivo final es seleccionar el clasificador que mejor se ajuste al mismo grupo de individuos. El método más común para la comparación de proporciones apareadas es el test de McNemar [68].

Sin pérdida de generalidad, consideremos el test de McNemar para la comparación de dos valores de sensibilidad dado un valor fijo de especificidad. El procedimiento es análogo para comparar especificidades dado un valor fijo de sensibilidad. El objetivo será comparar si los valores de sensibilidad que producen el clasificador 1 (Se_1) y el

clasificador 2 (Se_2) son significativamente diferentes:

$$H_0 : Se_1 = Se_2$$
$$H_1 : Se_1 \neq Se_2$$

Consideremos la tabla de contingencia 1.1, que representa las clasificaciones (positivo o negativo) de los clasificadores 1 y 2, para la muestra de observaciones pertenecientes a la clase 1. Nos interesa comparar si la diferencia entre la proporción $p_1 = \frac{a+b}{n_1}$ y la $p_2 = \frac{a+c}{n_1}$ es cero (hipótesis nula), esto es, $p_1 - p_2 = \frac{b-c}{n_1} \neq 0$.

	Clasificador 2		
Clasificador 1	Positivo	Negativo	Total
Positivo	a	b	$a+b$
Negativo	c	d	$c+d$
Total	$a+c$	$b+d$	$n_1 = a+b+c+d$

Tabla 1.1: Tabla de contingencia. Test de McNemar

El estadístico de contraste viene dado por:

$$Z = \frac{p_1 - p_2}{SE(p_1 - p_2)} = \frac{b-c}{\sqrt{b+c}} \sim N(0,1) \tag{1.28}$$

o, alternativamente:

$$\chi^2 = \frac{(b-c)^2}{b+c} \tag{1.29}$$

que sigue asintóticamente una distribución de chi-cuadrado con un grado de libertad. Si las frecuencias son pequeñas, puede producirse sesgo debido a la aproximación de una distribución discreta por una continua. En estos casos, puede utilizarse la corrección de Yates [107]: $\chi^2 = \frac{(|b-c|-1)^2}{b+c}$.

Capítulo 2

Métricas derivadas de la curva ROC

La curva ROC proporciona una visión global de la capacidad de un modelo de clasificación binaria en varios puntos de corte, pero no ofrece una medida específica para resumir su rendimiento. No obstante, pueden derivarse de ella métricas o medidas de evaluación, las cuales permiten cuantificar el rendimiento de manera más precisa. Algunas de estas medidas incluyen la sensibilidad, la especificidad y las tasas de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Estas métricas de evaluación, definidas en la Definición 3, proporcionan un valor específico que se determina mediante la elección de un punto de corte concreto, que permite asignar al individuo a su clase de pertenencia. Otras métricas de evaluación, como el valor predictivo positivo o valor predictivo negativo (Definición 3), también brindan una medida que evalúa la capacidad discriminatoria del modelo. Estas métricas optimizan las predicciones correctas de una clase o ambas, y su elección está determinada por el interés y el contexto del problema.

Además de las métricas presentadas en el capítulo anterior, existen otras métricas resumen derivadas de la curva ROC que han recibido atención en la literatura. Estas métricas son la conocida área bajo la curva ROC (AUC, siglas en inglés) o el índice de Youden, que son presentadas a continuación, así como sus variantes. En concreto, estas métricas (AUC e índice de Youden) se han utilizado para optimizar los modelos estimados en trabajos de la tesis.

2.1. Área bajo la curva ROC

El área bajo la curva ROC (AUC) es una de las medidas resumen más populares y ampliamente estudiadas en la literatura que derivan de la curva ROC. A diferencia de otras métricas, el AUC ofrece una medida de la bondad de ajuste global, abarcando todos los posibles puntos de corte.

Definición 6. *Dada la curva ROC (ROC(t)), definida en (1.2), el área bajo la curva*

ROC (AUC) se define como:

$$AUC = \int_0^1 ROC(t)dt \tag{2.1}$$

Observación 4. *El valor del AUC se sitúa en un rango de 0.5 a 1, donde valores más cercanos a 1 indican una mayor capacidad discriminativa. En los ejemplos presentados en la Figura 1.1, es evidente que el área bajo la curva ROC resultante para los casos extremos (Figura 1.1b y Figura 1.1c) sería de 0.5 y 1 (clasificador aleatorio y clasificador perfecto, respectivamente). Sin embargo, el caso más común es el que se muestra en la Figura 1.1a, donde el AUC toma valores intermedios en (0.5,1).*

Observación 5. *Es inmediato deducir que si la curva ROC para un clasificador A no se encuentra en ningún punto por debajo de la curva ROC del clasificador B ($ROC_A(t) \geq ROC_B(t)$), entonces las AUC correspondientes están ordenadas de la misma manera ($AUC_A \geq AUC_B$). Sin embargo, el recíproco no siempre es cierto; por ejemplo, ambas curvas pueden cruzarse y tener el mismo área bajo la curva.*

Bamber [10] demostró que el área bajo la curva ROC se puede interpretar como la probabilidad de que, dado un individuo sano y otro enfermo, el clasificador asigne un valor más alto al enfermo que al sano, tal y como se enuncia en la siguiente proposición.

Proposición 4. *Sean Y_0 e Y_1 variables independientes que representan los resultados del clasificador para la clase 0 (sanos) y la clase 1 (enfermos), respectivamente, entonces:*

$$AUC = P(Y_1 > Y_0) \tag{2.2}$$

Demostración. Asumimos por simplicidad que las funciones de distribución de Y_0 e Y_1, F_{Y_0} y F_{Y_1} respectivamente, son continuas. De la definición de la curva ROC (1.3), se tiene que:

$$AUC = \int_0^1 \mathrm{ROC}(t)dt = \int_0^1 \overline{F}_{Y_1}(\overline{F}_{Y_0}^{-1}(t))dt$$

donde $\overline{F}_{Y_k}$ denota la función de supervivencia, $k = 0, 1$. Usando el cambio de variable de t a $y = \overline{F}_{Y_0}^{-1}(t)$, la expresión queda de la siguiente manera:

$$\begin{aligned} AUC &= \int_{\infty}^{-\infty} \overline{F}_{Y_1}(y)d\overline{F}_{Y_0}(y) \\ &= \int_{-\infty}^{\infty} P(Y_1 > y)f_{Y_0}(y)dy \end{aligned}$$

donde f_{Y_0} es la función de densidad de Y_0. Finalmente, dada la independencia de las variables, se demuestra la proposición:

$$AUC = \int_{-\infty}^{\infty} P(Y_1 > y, Y_0 = y)dy = P(Y_1 > Y_0)$$

□

De la misma manera que la curva ROC, el AUC puede ser estimado mediante enfoques de estimación paramétrica y no paramétrica. A continuación, se exponen ambas estimaciones.

2.1.1. Estimación paramétrica

Supondremos ahora que los resultados del clasificador se distribuyen como $Y_0 \sim N(\mu_0, \sigma_0)$ e $Y_1 \sim N(\mu_1, \sigma_1)$, y considerando la independencia entre ellos, se tiene que $Y_1 - Y_0 \sim N(\mu_1 - \mu_0, \sqrt{\sigma_1^2 + \sigma_0^2})$ y, consecuentemente, $Z = \frac{(Y_1 - Y_0) - (\mu_1 - \mu_0)}{\sqrt{\sigma_1^2 + \sigma_0^2}} \sim N(0, 1)$.

Considerando la definición del AUC (2.2) y tomando en cuenta la suposición anterior, se obtiene la expresión paramétrica del AUC:

$$AUC = P(Y_1 > Y_0) = P(Y_1 - Y_0 > 0) \tag{2.3}$$

$$= P\left(Z > -\frac{(\mu_1 - \mu_0)}{\sqrt{\sigma_1^2 + \sigma_0^2}}\right) \tag{2.4}$$

$$= \Phi\left(\frac{\mu_1 - \mu_0}{\sqrt{\sigma_1^2 + \sigma_0^2}}\right) \tag{2.5}$$

$$= \Phi\left(\frac{a}{\sqrt{1 + b^2}}\right) \tag{2.6}$$

donde Φ es la función de distribución de la normal estándar, $a = \frac{\mu_1 - \mu_0}{\sigma_1}$ y $b = \frac{\sigma_0}{\sigma_1}$.

Por consiguiente, el AUC estimado vendrá dado por la estimación de los parámetros a y b: $\hat{a} = \frac{\hat{\mu}_1 - \hat{\mu}_0}{\hat{\sigma}_1}, \hat{b} = \frac{\hat{\sigma}_0}{\hat{\sigma}_1}, \widehat{AUC} = \Phi\left(\frac{\hat{a}}{\sqrt{1+\hat{b}^2}}\right)$. Esta estimación puede realizarse mediante el método de máxima verosimilitud.

Intervalos de confianza

Para obtener intervalos de confianza para el AUC, es necesario calcular su varianza. Liu y Schisterman [54] proponen la siguiente expresión que se deriva de aplicar el método delta [77]:

$$Var(\widehat{AUC}) = \frac{1}{\sigma_1^2 + \sigma_0^2} \Phi^2\left(\frac{\mu_1 - \mu_0}{\sqrt{\sigma_1^2 + \sigma_0^2}}\right) \left[Var(\hat{\mu}_0) + Var(\hat{\mu}_1) + \frac{(\mu_1 - \mu_0)^2}{4(\sigma_1^2 + \sigma_0^2)^4}(Var(\hat{\sigma}_0^2) + Var(\hat{\sigma}_1^2))\right] \tag{2.7}$$

2.1.2. Estimación no paramétrica

Cuando la asunción de normalidad no se cumple, es necesario recurrir a métodos no paramétricos para estimar el AUC.

Por definición (2.1), el AUC estimado puede obtenerse a partir de la curva ROC empírica (1.14) utilizando la regla del trapecio. Esta regla es un método de

integración que se basa en aproximar el área bajo la curva como suma de áreas de trapecios. Los límites de cada trapecio vendrán dados por cada punto de corte y su valor correspondiente en la curva escalonada. Dada una partición del intervalo (0,1) constituida por los puntos $\{t_0, t_1, .., t_{i-1}, t_i, ..., t_T\}$, el AUC estimado vendrá dado por la siguiente expresión:

$$\widehat{AUC} = \sum_{i=1}^{T} \frac{1}{2}(FPR_{t_i} - FPR_{t_{i-1}})(TPR_{t_i} + TPR_{t_{i-1}}) \tag{2.8}$$

donde FPR_{t_i} y TPR_{t_i} son las tasas de falsos y verdaderos positivos calculadas para el punto t_i.

Por otro lado, Bamber [10] proporciona una expresión para la estimación empírica del AUC, el cual es un estimador insesgado equivalente al estadístico U de Mann-Whitney [64, 41]:

$$\widehat{AUC} = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \Psi(y_{1j}, y_{0i}) \tag{2.9}$$

$$= \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \left[I(y_{1j} > y_{0i}) + \frac{1}{2} I(y_{1j} = y_{0i}) \right] \tag{2.10}$$

donde y_{ki} es la salida del clasificador para el individuo $i = 1, 2, \ldots, n_k$ de la clase $k = 0, 1$, I la función indicadora y

$$\Psi(x, y) = \begin{cases} 1 & si \quad x > y \\ \frac{1}{2} & si \quad x = y \\ 0 & si \quad x < y \end{cases} \tag{2.11}$$

Intervalos de confianza

Diferentes enfoques no paramétricos han sido propuestos para la estimación de la varianza del AUC y, consecuentemente, para la construcción de intervalos de confianza del AUC. En concreto, Hanley y McNeil [41], a partir de la igualdad (2.9), formulan la expresión del error estándar del AUC empírico como:

$$SE(\widehat{AUC}) = \sqrt{\frac{\widehat{AUC}(1 - \widehat{AUC}) + (n_1 - 1)(Q_1 - \widehat{AUC}^2) + (n_0 - 1)(Q_2 - \widehat{AUC}^2)}{n_0 n_1}} \tag{2.12}$$

siendo Q_1 la probabilidad de que dos individuos enfermos seleccionados aleatoriamente tengan un valor mayor del clasificador que un individuo sano, y Q_2 la probabilidad de que un individuo enfermo seleccionado al azar tenga un valor mayor del clasificador que dos individuos sanos seleccionados aleatoriamente.

Hanley y McNeil señalan que, para áreas del 80 % o más, el modelo exponencial negativo proporciona errores estándar ligeramente más conservadores que los otros modelos considerados. Bajo este supuesto, los valores de Q_1 y Q_2 pueden expresarse en función del AUC:

$$Q_1 = \frac{\widehat{AUC}}{2 - \widehat{AUC}} \tag{2.13}$$

$$Q_2 = \frac{2\widehat{AUC}^2}{1 + \widehat{AUC}} \tag{2.14}$$

A partir de las estimaciones de las expresiones anteriores se puede construir un intervalo de confianza para el AUC, así como determinar el tamaño de muestra necesario para conseguir una amplitud determinada [129].

Uno de los intervalos para el AUC más populares es el intervalo no paramétrico de DeLong [25], que se basa en la estimación de Mann-Whitney (2.9). El enfoque propuesto por DeLong et al. se fundamenta en el método de Sen [99] para obtener un estimador de la matriz de covarianza del AUC, que se basa en la teoría de los U-estadísticos. Definen como componentes estructurales las siguientes expresiones:

$$V_{10}^r(y_{1j}) = \frac{1}{n_0}\sum_{i=1}^{n_0}\Psi(y_{1j}^r, y_{0i}^r), \quad j = 1, 2, \ldots, n_1 \tag{2.15}$$

$$V_{01}^r(y_{0i}) = \frac{1}{n_1}\sum_{j=1}^{n_1}\Psi(y_{1j}^r, y_{0i}^r), \quad i = 1, 2, \ldots, n_0 \tag{2.16}$$

donde $\Psi(x, y)$ está definida en (2.11) y el subíndice r representa el r-ésimo modelo considerado. Este subíndice se utilizará específicamente para comparar las áreas bajo la curva ROC generadas por diferentes modelos (siguiente sección). La matriz de covarianzas estimada para el vector de parámetros estimados $\hat{\theta} = (\hat{\theta}^1, \ldots, \hat{\theta}^k)$, siendo $\hat{\theta}^k = \widehat{AUC_k}$ definido en (2.10), queda expresada como sigue:

$$\mathbf{S} = \frac{1}{n_1}\mathbf{S_{10}} + \frac{1}{n_0}\mathbf{S_{01}} \tag{2.17}$$

donde $\mathbf{S_{10}}$ y $\mathbf{S_{01}}$ son las matrices $k \times k$ donde el (r, s)-elemento viene dado, respectivamente, por:

$$s_{10}^{r,s} = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}[V_{10}^r(y_{1j}) - \hat{\theta}^r][V_{10}^s(y_{1j}) - \hat{\theta}^s] \tag{2.18}$$

$$s_{01}^{r,s} = \frac{1}{n_0 - 1}\sum_{i=1}^{n_0}[V_{01}^r(y_{0i}) - \hat{\theta}^r][V_{01}^s(y_{0i}) - \hat{\theta}^s] \tag{2.19}$$

donde $V_{10}^r(y_{1j})$ y $V_{01}^r(y_{0i})$ vienen dados por las expresiones (2.15)-(2.16). A partir de las

expresiones anteriores se deduce el estimador de la varianza de $\widehat{AUC}$:

$$\widehat{Var}(\widehat{AUC}) = \frac{s_{10}}{n_1} + \frac{s_{01}}{n_0} \tag{2.20}$$

$$= \frac{1}{n_1(n_1-1)} \sum_{j=1}^{n_1} (V_{10}(y_{1j}) - \widehat{AUC})^2 + \frac{1}{n_0(n_0-1)} \sum_{i=1}^{n_0} (V_{01}(y_{0i}) - \widehat{AUC})^2 \tag{2.21}$$

Considerando la estimación de la varianza (2.21), se puede construir un intervalo de confianza de tipo Wald para el AUC: $\hat{AUC} \pm z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{AUC})}$.

Otros autores, como Qin y Hotilovac [87], comparan varios métodos para la construcción de intervalos de confianza no paramétricos para el AUC, obteniendo diferentes resultados. Estos incluyen la prueba de Mann-Whitney [64], una transformación logit, el intervalo no paramétrico de DeLong [25], la verosimilitud empírica [88] e intervalos de confianza bootstrap. En concreto, los intervalos de confianza bootstrap representan la alternativa más frecuente dentro de los métodos empíricos.

2.1.3. Comparación de áreas bajo la curva ROC

La comparación del rendimiento entre dos modelos también se puede realizar mediante pruebas estadísticas de contraste, lo que nos permite seleccionar con mayor precisión el modelo con mejor rendimiento en caso de rechazar la hipótesis de igualdad. En esta sección, se presentan dos enfoques para la comparación de las áreas bajo la curva ROC, basados en los estudios de Hanley y McNeil [41] y DeLong et al. [25],

En este contexto, el objetivo será comparar las áreas bajo la curva ROC resultantes del modelo 1 (AUC_1) y del modelo 2 (AUC_2):

$$H_0 : AUC_1 = AUC_2$$
$$H_1 : AUC_1 \neq AUC_2$$

El estadístico de contraste propuesto por Hanley y McNeil [42] viene dado por:

$$Z = \frac{AUC_1 - AUC_2}{SE(AUC_1 - AUC_2)} \sim N(0,1) \tag{2.22}$$

siendo

$$SE(\widehat{AUC}_1 - \widehat{AUC}_2) = \sqrt{Var(\widehat{AUC}_1) + Var(\widehat{AUC}_2) - 2rSE(\widehat{AUC}_1)SE(\widehat{AUC}_2)} \tag{2.23}$$

donde r representa la correlación estimada entre $\widehat{AUC}_1$ y $\widehat{AUC}_2$. Es importante destacar que el último término es necesario para la comparación de áreas bajo la curva

ROC en casos apareados, es decir, cuando se analiza la capacidad predictiva de los clasificadores en el mismo conjunto de datos, que es lo común en la práctica.

El único parámetro desconocido de la expresión (2.23) es r, puesto que las expresiones de las varianzas del AUC estimado pueden ser sustituidas por las fórmulas presentadas anteriormente. Por ejemplo, se puede considerar la expresión para la varianza derivada de Hanley y McNeil, formulada en (2.12).

El parámetro r de correlación entre áreas se puede calcular de manera tradicional utilizando el coeficiente de correlación de Fisher/Pearson o el tau de Kendall [49] para datos ordinales. Hanley y McNeil [41] proponen calcular este coeficiente r tomando la media de los coeficientes calculados para cada subconjunto de datos de cada clase. En concreto, se define r como:

$$r = \frac{r_0 + r_1}{2} \tag{2.24}$$

donde r_0 representa el coeficiente de correlación obtenido al considerar los datos de la clase 0 y r_1 los de la clase 1.

Otra alternativa es utilizar las fórmulas derivadas en el estudio de DeLong et al. [25] y sustituirlas en la expresión (2.23), lo que resulta en el conocido test DeLong para la comparación de áreas bajo la curva ROC, ampliamente utilizado en problemas de clasificación. Específicamente, se considera la fórmula de la varianza dada en (2.21) y, de manera similar, se obtiene la expresión de la covarianza utilizando las fórmulas (2.18)-(2.19) que definen los elementos de la matriz de covarianza estimada (2.17).

2.1.4. Área parcial bajo la curva ROC

Aunque las pruebas de comparación de áreas bajo la curva ROC ayudan a elegir el mejor modelo según la discriminación medida por el AUC, no garantizan que las curvas ROC tengan formas similares si no se rechaza la hipótesis nula de igualdad de áreas. De hecho, es posible encontrar dos pruebas con curvas ROC que presenten formas muy diferentes, pero cuyas AUC sean prácticamente idénticas. Esto se debe a que el AUC evalúa la discriminación del modelo desde un enfoque global. Sin embargo, el objetivo de comparación puede que se focalice en una región concreta del espacio de curvas ROC. En concreto, si el interés se centra en un rango específico de la tasa de falsos positivos (FPR), la métrica adecuada sería el área parcial del AUC.

Definición 7. *Dado un intervalo (a,b), donde* $0 \leq a < b \leq 1$*, se define el área parcial bajo la curva ROC (pAUC, por sus siglas en inglés) en dicho intervalo, como el área encerrada entre la curva ROC, el eje horizontal y las líneas verticales en las coordenadas* $x{=}a$ *y* $x{=}b$*:*

$$pAUC_{(a,b)} = \int_a^b ROC(t)dt \qquad 0 \leq a < b \leq 1 \tag{2.25}$$

La estimación del área parcial bajo la curva ROC puede realizarse, por definición, utilizando las estimaciones para las curvas ROC ($\widehat{ROC(t)}$) mediante métodos paramétricos o método Kernel:

$$\widehat{pAUC}_{(a,b)} = \int_a^b \widehat{ROC(t)}dt \quad 0 \leq a < b \leq 1 \tag{2.26}$$

Para el caso de estimaciones no paramétricas de la curva ROC escalonadas (no suavizadas), se puede aplicar el método trapezoidal, similar a (2.8).

2.2. Índice de Youden

El índice de Youden es una métrica formulada por William J. Youden [125] que ha sido ampliamente utilizada en estudios clínicos, aunque no ha recibido tanta atención en la literatura como el AUC. Ambas medidas son útiles para evaluar modelos de clasificación pero se enfocan en diferentes aspectos de la evaluación del rendimiento del modelo. Mientras que el AUC evalúa la capacidad discriminativa general del modelo considerando todos los puntos de corte, el índice de Youden proporciona un valor específico de rendimiento del modelo a partir de la sensibilidad y la especificidad. Esto implica elegir un punto de corte concreto que maximice el índice de Youden. La ventaja de esto en comparación con el AUC es que se proporciona un punto de corte que ofrece al clínico un valor de decisión que permite interpretar y clasificar a los pacientes entre la clase positiva (enfermos) y la negativa (sanos).

Definición 8. *El cálculo del índice de Youden se basa en la diferencia entre la tasa de verdaderos positivos y la tasa de falsos positivos tal que:*

$$J(c) = TPR(c) - FPR(c) = Sensibilidad(c) + Especificidad(c) - 1 = F_{Y_0}(c) - F_{Y_1}(c) \tag{2.27}$$

donde c denota el punto de corte y F_{Y_k} la función de distribución de la variable respuesta Y_k, $k = 0, 1$.

El índice de Youden óptimo J_{max} se alcanza maximizando la diferencia anterior (2.27):

$$J_{max} = max_c\{Sensibilidad(c) + Especificidad(c) - 1\} \tag{2.28}$$

$$= max_c\{F_{Y_0}(c) - F_{Y_1}(c)\} \tag{2.29}$$

Observación 6. *El índice de Youden varía entre 0 y 1, donde valores cercanos a 0 indican una capacidad discriminatoria baja y valores cercanos a 1 indican una capacidad discriminatoria alta. En concreto, un valor de 1 representa una clasificación perfecta, lo que significa que la pueba no produce falsos positivos ni falsos negativos (Sensibilidad=Especificidad=1; o equivalentemente: TPR=1, FPR=0).*

El índice de Youden, al igual que el AUC, es una métrica derivada de la curva ROC y, por lo tanto, puede estimarse utilizando enfoques paramétricos y no paramétricos (Fluss [32]). Estos enfoques se describen a continuación.

2.2.1. Estimación paramétrica

El enfoque paramétrico para estimar el índice de Youden bajo normalidad se basa en los resultados presentados por Schisterman y Perkins [98].

Asumiendo que los resultados del clasificador se distribuyen como $Y_0 \sim N(\mu_0, \sigma_0)$ y $Y_1 \sim N(\mu_1, \sigma_1)$ con $\sigma_0^2 \neq \sigma_1^2$, la función que proporciona el índice de Youden se expresa como:

$$J(c) = TPR(c) - FPR(c) = \Phi\left(\frac{c-\mu_0}{\sigma_0}\right) - \Phi\left(\frac{c-\mu_1}{\sigma_1}\right) \tag{2.30}$$

donde Φ es la función de distribución de la normal estándar.

Para maximizar la expresión anterior (2.30), se deriva con respecto a c y se establece la derivada igual a cero. Esto conduce a la ecuación cuadrática resultante que proporciona el punto de corte óptimo:

$$c^* = \frac{(\mu_1\sigma_0^2 - \mu_0\sigma_1^2) - \sigma_0\sigma_1\sqrt{(\mu_0-\mu_1)^2 + (\sigma_0^2-\sigma_1^2)\log(\frac{\sigma_0^2}{\sigma_1^2})}}{\sigma_0^2 - \sigma_1^2} \tag{2.31}$$

Sustituyendo (2.31) en la ecuación (2.30), se obtiene el índice de Youden máximo. Si se asumiese igualdad de varianzas ($\sigma_0^2 = \sigma_1^2$), el punto de corte óptimo sería:

$$c^* = \frac{(\mu_0 + \mu_1)}{2} \tag{2.32}$$

y el índice de Youden máximo:

$$J(c^*) = 2\Phi\left(\frac{\mu_1 - \mu_0}{2\sqrt{\sigma_1^2}}\right) - 1 \tag{2.33}$$

Estas formulaciones son también válidas bajo transformaciones tipo Box-Cox.

Schisterman y Perkins [98] también abordan la estimación de la varianza del índice de Youden para la creación de intervalos de confianza utilizando el método delta. Como alternativa, proponen también la estimación aplicando diversas técnicas de bootstrap.

2.2.2. Estimación no paramétrica

El enfoque no paramétrico más sencillo se basa en considerar las funciones de distribución empíricas, formuladas en (1.13). Sustituyendo en (2.27) se obtiene la estimación del índice de Youden:

$$\hat{J}(c^*) = \hat{F}_{Y_0}(c^*) - \hat{F}_{Y_1}(c^*) \tag{2.34}$$

$$= \frac{1}{n_0}\sum_{i=1}^{n_0} I(y_{0i} \leq c^*) - \frac{1}{n_1}\sum_{i=1}^{n_1} I(y_{1i} \leq c^*) \tag{2.35}$$

siendo $c^* \in \{y_{01}, \ldots, y_{0n_0}, y_{11}, \ldots, y_{1n_1}\}$ el punto de corte óptimo que maximiza la expresión.

Otro enfoque no paramétrico para la estimación del índice de Youden es el método tipo Kernel. En este caso, el índice de Youden estimado máximo tiene la siguiente expresión (considerando el núcleo gaussiano):

$$\hat{J}(c^*) = \hat{F}_{Y_0}(c^*) - \hat{F}_{Y_1}(c^*) \tag{2.36}$$

$$= \frac{1}{n_0} \sum_{i=1}^{n_0} \Phi\left(\frac{c^* - y_{0i}}{h_{Y_0}}\right) - \frac{1}{n_1} \sum_{i=1}^{n_1} \Phi\left(\frac{c^* - y_{1i}}{h_{Y_1}}\right) \tag{2.37}$$

donde c^* representa el punto de corte óptimo, Φ es la función de distribución de una normal estándar y h_{Y_k} el ancho de banda, cuya elección común se determina según la ecuación (1.19). En un estudio comparativo realizado por Faraggi y Reiser [30] sobre diferentes métodos de selección de anchos de banda, concluyeron que procedimientos más complejos no generaban mejoras significativas.

En cuanto a la estimación de los intervalos de confianza para el índice de Youden, Zhou y Qin [127] proponen intervalos de confianza no paramétricos, los cuales recomiendan cuando se carece de información sobre las distribuciones de las poblaciones enfermas y no enfermas. Los intervalos de confianza propuestos se basan en la estimación ajustada de Agresti y Coull [1] para una proporción binomial y la estimación bootstrap de la varianza del índice de Youden empírico. Siguiendo otro enfoque, Shan [100] propone dos intervalos de confianza utilizando el método de Wilson-score [120].

2.2.3. Adaptaciones del índice de Youden

Ciertas modificaciones del índice de Youden se han propuesto con el fin de adaptarlo a diferentes contextos y necesidades.

El índice de Youden, definido en (2.27), ofrece un equilibrio optimizado entre sensibilidad y especificidad. Sin embargo, en ciertas situaciones clínicas, es posible que se desee otorgar mayor importancia a la sensibilidad en comparación con la especificidad, o viceversa. También puede haber interés en optimizar la métrica dentro de un rango específico de sensibilidades y especificidades en el espacio de curvas ROC. Es así como surgen el denominado índice de Youden ponderado e índice de Youden parcial, que son presentados a continuación.

Índice de Youden ponderado

El índice de Youden ponderado es una métrica adaptada del índice de Youden que combina la sensibilidad y la especificidad de una prueba diagnóstica teniendo en cuenta

a su vez la importancia relativa de cada una de estas medidas. Autores, como Rückera y Schumache [93] y Li et al. [52], han utilizado y definido esta métrica. A continuación se presentan las definiciones que proporcionan estos autores.

Rückera y Schumache [93] utilizan la versión ponderada del índice de Youden que se expresa de la siguiente manera:

$$J_w(c) = w \cdot Sensibilidad(c) + (1 - w) \cdot Especificidad(c) \tag{2.38}$$

donde w, $0 < w < 1$ es el párametro que actúa en la ponderación (peso) y c el punto de corte.

Más tarde, Li et al. [52] definen el índice de Youden ponderado de la siguiente manera:

$$J_w(c) = 2(w \cdot Sensibilidad(c) + (1 - w) \cdot Especificidad(c)) - 1 \tag{2.39}$$

con $0 \leq w \leq 1$. Esta expresión es equivalente al índice de Youden (2.27) cuando $w = 0{,}5$.

Índice de Youden parcial

Recientemente, Li et al. [51] definieron el índice de Youden parcial, con una motivación similar a la del AUC parcial (section 2.1.4), cuya definición se introduce a continuación.

Definición 9. *Dado el intervalo (a,b) con* $0 \leq a < b \leq 1$*, se define el índice de Youden parcial máximo en dicho intervalo como:*

$$\begin{aligned} J_{a,b}(c^*) &= max_{a \leq c \leq b}\{Sensibilidad(c) + Especificidad(c) - 1\} && (2.40)\\ &= max_{a \leq c \leq b}\{F_{Y_0}(c) - F_{Y_1}(c)\} && (2.41)\\ &= Sensibilidad(c^*) + Especificidad(c^*) - 1 && (2.42)\\ &= F_{Y_0}(c^*) - F_{Y_1}(c^*) && (2.43) \end{aligned}$$

donde c^* *es el punto de corte óptimo y* (a, b) *el intervalo de puntos de corte determinados a partir de un rango de valores de sensibilidad y especificidad deseados.*

A diferencia del índice de Youden, optimizado en todo el rango de valores de sensibilidad y especificidad, el índice de Youden parcial se focaliza en una región concreta del espacio ROC.

En resumen, el índice de Youden y sus variantes tienen una relevancia tanto estadística como clínica, puesto que ofrecen un valor de la capacidad del clasificador a

la vez que proporcionan un criterio para seleccionar un punto de corte óptimo, como aquel que maximiza las expresiones anteriores. El principal interés de la tesis se ha focalizado en la optimización de modelos de clasificación bajo optimalidad del índice de Youden.

La selección del punto de corte óptimo es un aspecto fundamental en los modelos de clasificación y puede variar según el contexto y objetivo del estudio clínico. En el próximo capítulo se presentan diferentes enfoques para la selección del punto de corte óptimo.

Capítulo 3

Selección del punto de corte óptimo

En los capítulos previos, se han presentado las estimaciones de la curva ROC y las métricas derivadas que proporcionan una evaluación del rendimiento de los modelos de clasificación. En la práctica general, esta evaluación cuantitativa del rendimiento del modelo debe llevar consigo la determinación y aplicación de un punto de corte específico que permita trasladar esta evaluación a decisiones prácticas y útiles. En el contexto de la biomedicina, esto implica la necesidad de establecer un umbral que oriente la decisión de clasificar a un paciente como positivo para la enfermedad o condición, basándose en la predicción continua del modelo.

La elección de este punto de corte óptimo dependerá del contexto y de los objetivos específicos del estudio clínico, lo que implica generalmente decidir qué métrica de evaluación o combinación se busca optimizar. Puede ser de interés identificar, por ejemplo, la mayor cantidad posible de casos positivos (enfermos) o, por otro lado, ser más preciso en las clasificaciones de individuos sanos o enfermos. En consecuencia, la selección de un punto de corte óptimo se convierte en un aspecto crucial para garantizar la aplicación efectiva de los modelos de clasificación en entornos clínicos o prácticos.

En este capítulo se introducen diferentes métodos y enfoques para la selección del punto de corte óptimo, ofreciendo una herramienta valiosa para la toma de decisiones informadas.

3.1. Criterios basados en la sensibilidad o especificidad

Un posible criterio para la selección del punto de corte óptimo c^* podría ser aquel que maximiza la sensibilidad o la especificidad. Sin embargo, por lo general, no es adecuado optimizar la sensibilidad y la especificidad de forma aislada, ya que maximizar la sensibilidad podría resultar en una disminución de la especificidad, y viceversa. Por ejemplo, en circunstancias extremas donde el foco fuese la maximización

de la sensibilidad, podría resultar en la clasificación de todos los pacientes como pertenecientes a la clase positiva (enfermos), resultando en una sensibilidad máxima de 1. Sin embargo, los falsos positivos suelen estar vinculados con daños para el paciente o consecuencias económicas importantes, por lo que generalmente no se considera un buen criterio. Habitualmente se busca criterios que equilibren ambas métricas.

Un enfoque alternativo sería considerar el punto de corte c^* donde la sensibilidad y especificidad son equivalentes $Se(c^*) \approx Sp(c^*)$ o:

$$c^* = \arg\min_c \{Se(c) - Sp(c)\} \tag{3.1}$$

Este punto se conoce como punto de simetría o equivalencia y representa matemáticamente el punto de intersección entre la curva ROC y la línea $y = 1 - x$. Puede interpretarse como el punto que maximiza simultáneamente ambos tipos de clasificaciones correctas: verdaderos positivos y verdaderos negativos.

3.1.1. Criterio basado en el índice de Youden

El índice de Youden, definido en (2.27), es una métrica que busca maximizar la suma de la sensibilidad y la especificidad, asignándoles el mismo peso. Este índice se emplea ampliamente para seleccionar el punto de corte óptimo cuando se busca equilibrar estas dos métricas. El punto de corte óptimo c^* asociado a este índice corresponde al punto en la curva ROC donde la distancia a la diagonal $TPR = FPR$ alcanza su valor máximo. La interpretación intuitiva implica identificar el punto en la curva que está más distante del azar o, en otras palabras, más alejado de ninguna discriminación.

En ausencia de un criterio claro para asignar valores más altos a la sensibilidad o la especificidad, el índice de Youden se presenta como una métrica idónea, proporcionando un equilibrio óptimo entre ambas métricas y considerando de igual manera tanto los verdaderos positivos como los verdaderos negativos. Esta característica confiere al índice de Youden una mayor robustez frente a desbalances de clases en comparación con otras medidas como la tasa de verdaderos positivos, permitiendo así una evaluación equitativa del rendimiento del modelo en ambas direcciones.

En conclusión, el índice de Youden no solo representa una métrica de rendimiento balanceada, sino que también sirve como criterio para la selección del punto de corte óptimo en la curva ROC. Es la métrica principal en la que se centra el trabajo de esta tesis. Una introducción a su definición y estimación ha sido detallada en el capítulo anterior (sección 2.2).

3.1.2. Punto de la curva ROC más cercano al (0,1)

Otra opción comúnmente utilizada para la selección del punto de corte óptimo es elegir aquel que esté más próximo al punto ideal de la curva ROC, esto es, $(0, 1)$. En otras palabras, define como punto de corte óptimo aquel que se encuentra más cerca de la discriminación perfecta entre individuos de las clases 0 y 1. Matemáticamente, este punto de corte c^* satisface la siguiente ecuación:

$$c^* = \arg\min_c \{\sqrt{(1 - TPR(c))^2 + FPR(c)^2}\} \tag{3.2}$$

Tanto el criterio basado en el punto más cercano a (0,1) como el índice de Youden son ampliamente utilizados y pueden proporcionar puntos de corte óptimos que pueden coincidir o diferir, según el análisis realizado por Perkins y Schisterman [84]. Sin embargo, los autores respaldan la preferencia por el uso del índice de Youden, resaltando tanto su interpretación matemático-clínica como su respaldo y trabajo en la literatura.

3.1.3. Coeficiente de correlación de Matthews

Una alternativa adicional para seleccionar el punto de corte óptimo es maximizar el coeficiente de correlación de Matthews [67], una medida estadística utilizada para evaluar el rendimiento de un clasificador binario, aunque su uso es menos común en comparación con los métodos mencionados anteriormente.

Definición 10. *El coeficiente de Matthews (MCC, siglas en inglés) para un punto de corte c se define de la siguiente manera:*

$$MCC(c) = \frac{VP \cdot VN - FP \cdot FN}{\sqrt{(VP + FN)(VN + FP)(VP + FP)(VN + FN)}} \tag{3.3}$$

Observación 7. *La correlación de Matthews toma valores en el rango de -1 a 1, donde 1 representa una clasificación perfecta y 0 sugiere que la clasificación es similar a la de un modelo aleatorio.*

3.1.4. Índice DOR

El Índice de Odds de Diagnóstico (DOR, siglas en inglés), introducido por Glas et al. [35], es una medida que evalúa la utilidad diagnóstica de una prueba.

Definición 11. *El índice DOR para un punto de corte c se define de la siguiente manera:*

$$DOR(c) = \frac{VP/FN}{FP/VN} = \frac{Se(c)/(1 - Se(c))}{(1 - Spe(c))/Spe(c)} \tag{3.4}$$

donde Se y Spe denotan la sensibilidad y especificidad, respectivamente.

Observación 8. *El índice de Odds de Diagnóstico (DOR) puede oscilar entre cero e infinito. Un DOR más alto sugiere una mejor capacidad diagnóstica de la prueba. Cuando el DOR es igual a 1, indica que la prueba carece de capacidad discriminativa, y valores más bajos podrían interpretarse como una clasificación peor que el azar. Con este enfoque, el punto de corte óptimo sería aquel que maximiza el índice DOR.*

A pesar de ser un índice más reciente que el índice de Youden, su impacto ha sido más limitado. En concreto, Böhning et al. [14] desaconsejan su uso, ya que demuestran que este criterio podría llevar a la elección de valores de punto de corte en el límite del rango de parámetros del valor de salida del modelo. Los autores indican que el índice de Youden es preferible.

3.1.5. Probabilidad de concordancia

Liu [57] introduce la probabilidad de concordancia como un enfoque para seleccionar el punto de corte óptimo en la curva ROC.

Definición 12. *Dado un punto de corte c determinado, se define la probabilidad de concordancia (PC) como el producto de la sensibilidad (Se) y la especificidad (Spe):*

$$PC(c) = Se(c) \times Spe(c) = TPR(c)(1 - FPR(c)) \tag{3.5}$$

El punto de corte óptimo según este criterio será aquel que maximice dicha expresión (3.5). La probabilidad de concordancia, al igual que el índice de Youden, tiene una interpretación intuitiva vinculada a la curva ROC. Se puede entender como el área de un rectángulo de lados $Se(c)$ y $Spe(c)$. Como el área máxima de un rectángulo corresponde a un cuadrado, es fácil deducir que este criterio favorece la equidad en Se y Spe.

Autores como Liu [57], Rota y Antolini [91] y Unal [111] presentan estudios que comparan los métodos presentados anteriormente para la selección del punto de corte óptimo.

En conclusión, aunque la elección entre los distintos criterios depende del objetivo específico del problema, en ausencia de consenso o si se busca ponderar de manera equitativa la sensibilidad y la especificidad, el índice de Youden es uno de los métodos más utilizados y recomendados en la literatura. Autores como Perkins y Schisterman [84], así como Böhning et al. [14], lo respaldan como el criterio más apropiado, prefiriéndolo sobre otros enfoques debido a su significado clínico y su amplio uso, proporcionando además una medida del rendimiento del modelo.

3.2. Enfoques basados en costes

En determinadas circunstancias, puede resultar más conveniente o necesario priorizar la clasificación precisa de una clase sobre otra. En estos casos, los enfoques previamente mencionados, como el índice de Youden, no ofrecerían un criterio adecuado para la selección del punto de corte óptimo, ya que no tienen en cuenta la prevalencia de la enfermedad al otorgar igual peso a la sensibilidad y la especificidad. A diferencia de los métodos anteriores, los criterios basados en la metodología coste-beneficio tienen en cuenta la prevalencia de la enfermedad y los costes relacionados con la clasificación incorrecta de los pacientes, ya sea diagnosticándolos como sanos cuando están enfermos o viceversa [133]. Estos enfoques permiten asignar diferentes pesos o importancias a los errores cometidos, lo cual es especialmente valioso en contextos clínicos donde el coste de diagnosticar incorrectamente a un paciente con una enfermedad puede ser diferente al coste de no diagnosticar a un paciente que realmente tiene la enfermedad. Por tanto, estos métodos resultan particularmente útiles cuando los errores de clasificación tienen consecuencias significativas. A continuación, se presentan enfoques basados en costes para la selección del punto de corte.

3.2.1. Método coste-beneficio

Se define el coste esperado a través de la siguiente expresión [69]:

$$C = C_0 + P(VP)C_{VP} + P(FP)C_{FP} + P(VN)C_{VN} + P(FN)C_{FN} \tag{3.6}$$

donde C_0 indica el coste base y $C_{VP}, C_{FP}, C_{VN}, C_{FN}$ los costes de las consecuencias de la decisión asociados a los VP, FP, VN y FN, respectivamente. $P(VP), P(FP), P(VN), P(FN)$ son las probabilidades de obtener un VP, FP, VN y FN, respectivamente:

$$P(VP) = P(positivo) \times Se = p \times TPR \tag{3.7}$$

$$P(FP) = P(negativo) \times (1 - Spe) = (1 - p) \times FPR \tag{3.8}$$

$$P(VN) = P(negativo) \times Spe = (1 - p) \times TNR \tag{3.9}$$

$$P(FN) = P(positivo) \times (1 - Se) = p \times FNR \tag{3.10}$$

siendo p la prevalencia de la enfermedad, Se sensibilidad y Sp especificidad.

Lema 1. *Sea una curva ROC asociada a un diagnóstico sobre un evento, el punto de corte óptimo que minimiza el coste asociado a los errores de clasificación (3.6) es aquel correspondiente al punto de la curva ROC cuya recta tangente tiene por pendiente:*

$$m = \frac{costes\ falsos\ positivos \times (1 - p)}{costes\ falsos\ negativos \times p}$$

Demostración. Dada la curva ROC, $x = FPR$ y $f(x) = TPR$, sustituyendo en (3.6), obtenemos:

$$C = C_0 + (1-p)C_{VN} + pC_{FN} + p(C_{VP} - C_{FN})f(x) + (1-p)(C_{FP} - C_{VN})x$$

Para hallar el mínimo de esta función de coste, tomamos la derivada e igualamos a cero. Despejando $f'(x)$ y considerando que los costes de los aciertos son nulos, nos queda:

$$f'(x) = \frac{(1-p)C_{FP}}{pC_{FN}}$$

Por tanto, el valor mínimo de la función de costes (3.6) se alcanzará en el punto que tenga por pendiente $f'(x)$, quedando demostrado el lema. □

En la práctica, para encontrar el punto de corte óptimo según este criterio, se puede trazar una recta desde el punto ideal (0,1) con pendiente m y desplazarla paralelamente hacia la curva ROC estimada. El punto de intersección de esta recta con la curva se asociaría con el punto de corte óptimo.

Por otro lado, asumiendo binormalidad de la curva ROC ($Y_0 \sim N(\mu_0, \sigma_0)$ y $Y_1 \sim N(\mu_1, \sigma_1)$), Somoza y Mossman [106] proporcionan una expresión que permite determinar el punto de corte óptimo. Sean los parámetros $a = \frac{\mu_1 - \mu_0}{\sigma_1}$ y $b = \frac{\sigma_0}{\sigma_1}$, las coordenadas de la curva ROC (FPR, TPR) que determinan el punto de corte óptimo para $b = 1$ vienen dadas por las siguientes expresiones:

$$FPR = \Phi\left(-\frac{a}{2} - \frac{ln(N)}{a}\right) \tag{3.11}$$

$$TPR = \Phi\left(-\frac{1}{2} - \frac{ln(N)}{a}\right) \tag{3.12}$$

$$\tag{3.13}$$

donde Φ es la función de distribución de la normal estándar y N el número de individuos de la muestra. Para $b \neq 1$, se tienen las siguientes expresiones:

$$FPR = \Phi\left(\frac{ab - \sqrt{a^2 - 2(1-b^2)ln(\frac{N}{b})}}{1-b^2}\right) \tag{3.14}$$

$$TPR = \Phi\left(\frac{a - b\sqrt{a^2 - (1-b^2)ln(\frac{N}{b})}}{1-b^2}\right) \tag{3.15}$$

$$\tag{3.16}$$

donde N es el número de individuos de la muestra y Φ la función de distribución de la normal estándar.

3.2.2. Término de costes por clasificación incorrecta

Otro criterio de selección de punto de corte óptimo basado en los costes de clasificaciones incorrectas es el del término de costes por clasificación incorrecta (MCT, siglas en inglés), definido por:

$$MCT(c) = \frac{C_{FN}}{C_{FP}} p(1 - Se(c)) + (1 - p)(1 - Spe(c)) \tag{3.17}$$

El punto de corte óptimo será aquel que minimice esta expresión (3.17) [12].

3.2.3. Punto de simetría generalizado

López-Ratón et al. [59, 60] se centraron en el punto de simetría para la determinación óptima del punto de corte. Los autores introducen el punto de simetría generalizado, que generaliza el punto de simetría (3.1) al incorporar los costes asociados a las clasificaciones erróneas C_{FP} y C_{FN}. El punto de simetría generalizado c_{GS} es aquel que satisface la siguiente ecuación:

$$\frac{C_{FP}}{C_{FP}}(1 - Spe(c_{GS})) = (1 - Se(c_{GS})) \tag{3.18}$$

3.2.4. Índice de Youden generalizado

El índice de Youden generalizado (GYI, siglas en inglés) surge como una extensión del índice de Youden, incorporando los costes de las clasificaciones incorrectas del diagnóstico, formulado de la siguiente forma (Lopez-Ratón [58]):

$$GYI(c) = Se(c) + \frac{1 - p}{p} \frac{C_{FN}}{C_{FP}} Sp(c) - 1 \tag{3.19}$$

El punto de corte óptimo será aquel que maximize dicho índice.

Observación 9. *Es evidente que el índice de Youden generalizado es equivalente al índice de Youden (definido en (2.27)) cuando* $\frac{1-p}{p}\frac{C_{FN}}{C_{FP}} = 1$. *Un ejemplo que cumple con esta equivalencia es cuando la prevalencia es* $p = 0{,}5$ *y los costes son iguales a 1.*

A pesar de que estos enfoques tengan en cuenta los costes asociados a clasificaciones incorrectas y permita asignar diferente importancia a los errores cometidos, la determinación precisa de estos costes puede resultar difícil de cuantificar en la práctica. Por esta razón, seleccionar el punto de corte óptimo basado en estos criterios no es la elección más frecuente.

3.3. Enfoques basados en métodos gráficos

Para determinar el punto de corte óptimo, se puede adoptar otra perspectiva mediante el uso de representaciones gráficas. Estos métodos se basan en visualizar las implicaciones de seleccionar un punto de corte específico, considerando tanto los beneficios como los perjuicios, ya sea en términos de individuos correctamente clasificados ($VN + VP$) o incorrectamente clasificados ($FN + FP$). En el ámbito biomédico, estos enfoques ofrecen a los profesionales clínicos una herramienta gráfica para la toma de decisiones informadas, facilitando la elección del punto de corte según los objetivos clínicos y teniendo en cuenta las consecuencias asociadas.

Los enfoques gráficos basados en funciones de densidad buscan identificar el punto de corte óptimo que maximice la discriminación entre individuos de las clases 0 y 1. Este proceso implica la representación visual de las funciones de densidad de ambos grupos. El objetivo es seleccionar el punto de corte a partir del gráfico que maximice la distancia entre las funciones de densidad, logrando una separación efectiva entre ambas clases.

Esta idea ya se había introducido en el Capítulo 1, con la Figura 1.1a, que muestra las funciones de densidad para una población enferma y sana. En ella, se selecciona el punto de corte que pondera de igual manera a los enfermos y a los sanos. Sin embargo, se podría seleccionar otro criterio para la elección del punto de corte. La representación gráfica tiene la ventaja de informar visualmente sobre cómo cambiaría la proporción de individuos correctamente clasificados (VN + VP) o incorrectamente clasificados (FN + FP) en función del punto de corte escogido. De esta manera, se podría ajustar este punto de corte para obtener un porcentaje más alto o más bajo de falsos negativos o falsos positivos, por ejemplo. Por tanto, suponiendo un clasificador que proporciona unos valores numéricos que corresponden a la probabilidad de pertenencia a la clase 1 (enferma), si conocemos las funciones de densidad de la población enferma y sana, obtendríamos un gráfico similar al de la Figura 1.1a donde se podría visualizar el efecto de seleccionar diferentes puntos de corte (probabilidades predichas por el clasificador) y comprender su efecto en términos de bien y mal clasificados para cada clase. La elección del punto de corte está condicionada por el modelo de clasificación y el propósito o utilidad del contexto.

Esta estrategia no solo simplifica la identificación visual del punto de corte, sino que también proporciona una comprensión intuitiva de la distribución de los datos en relación con la decisión de clasificación. De esta manera, conocer las funciones de densidad de cada clase nos posibilita obtener, a través de este enfoque, una perspectiva clara sobre la capacidad de discriminación del clasificador y el punto de corte óptimo,

de acuerdo con los requisitos específicos del objetivo en consideración. Aunque en la práctica las funciones de densidad raramente son conocidas, es posible estimarlas mediante la aproximación de funciones tipo kernel (1.15). Además, se sabe que, con un número suficiente de datos, las estimaciones proporcionadas se aproximan a las funciones de densidad poblacionales [104].

Autores como Borque et al. [15] y Rubio-Briones et al. [92] emplean este enfoque gráfico para la selección de puntos de corte en el contexto del cáncer de próstata. El estudio de Borque et al. se centra en la estadificación de este cáncer, una tarea crucial para determinar si el tumor está confinado dentro de la próstata. Los autores construyen curvas de densidad de probabilidades para evaluar diversos clasificadores con este propósito, lo que les permite elegir el punto de corte apropiado para distinguir entre grupos de alto y bajo riesgo. Por otro lado, en el estudio de Rubio-Briones et al., validan la utilidad del modelo para predecir el cáncer de próstata en la biopsia inicial, basando la selección del punto de corte óptimo en el estudio de las funciones de densidad asociadas al clasificador.

Software

Aunque existen varios programas estadísticos para analizar la curva ROC, como IBM SPSS, R destaca por ser un entorno de software libre y de código abierto. R ofrece todas las funciones necesarias para el análisis de curvas ROC a través de paquetes como ROCR [105], pROC [89] y OptimalCutpoints [61]. Estos paquetes proporcionan opciones para seleccionar el punto de corte óptimo, y en particular, OptimalCutpoints ofrece una amplia gama de criterios para esta selección que incluyen, además de los criterios previamente mencionados, otros relacionados como los basados en los valores predictivos. El software R se ha usado en los principales trabajos de la tesis debido a sus ventajas y utilidades. Asimismo, se han desarrollado librerías en R que están a libre disposición para la comunidad científica a través de repositorios de acceso público, cuya información se detalla en los siguientes capítulos. En [73] se puede consultar más información sobre el software estadístico disponible para el análisis de curvas ROC.

Capítulo 4

Modelos de clasificación

En este capítulo se presentan los algoritmos y métodos de clasificación que sirven como base para la construcción de los modelos presentados en los artículos de esta tesis. En concreto, destacar los modelos lineales optimizados por el AUC o criterios derivados de la curva ROC, que constituyen el fundamento de los algoritmos propuestos. Aunque el capítulo relaciona alguno de los conceptos introducidos con los artículos que constituyen el compendio, una descripción más detallada de estos se presenta de manera independiente en cada capítulo dedicado a ellos.

4.1. Modelos lineales bajo optimización del AUC

En el ámbito de la biomedicina, es habitual en la práctica clínica recopilar información sobre varios biomarcadores para el diagnóstico de enfermedades. La combinación de estos biomarcadores en un solo indicador es una práctica común, dado que suele proporcionar diagnósticos más efectivos que cada biomarcador por separado.

Los métodos de combinación lineal han sido ampliamente desarrollados y aplicados en este contexto de clasificación binaria, preferidos por su simplicidad y facilidad de interpretación. En ocasiones, estos métodos ofrecen un equilibrio adecuado entre estas características y su rendimiento. La combinación lineal de biomarcadores continuos genera un valor continuo, y la dicotomización de este valor resulta crucial, ya que proporciona al clínico una regla de clasificación relacionada con la enfermedad, sobre la cual puede fundamentar sus decisiones clínicas. En el capítulo 3 se han presentado diversos enfoques para la selección del mejor punto de corte.

El objetivo final sería formular el modelo lineal con una mayor capacidad de discriminación. La evaluación del rendimiento de esta combinación lineal (marcador diagnóstico) se puede realizar a través de medidas derivadas de la curva ROC, como pares de sensibilidad y especificidad, el área o el área parcial bajo la curva ROC, y el índice de Youden (capítulo 2). Del enfoque derivado de la selección de modelos óptimos

bajo criterio de parámetros de la curva ROC han surgido algoritmos que estiman los parámetros del modelo lineal con este tipo de objetivos. En concreto, la optimización del AUC es el que ha tenido un mayor número de propuestas y desarrollo.

4.1.1. Estimación paramétrica

El análisis discriminante fue formulado por Fisher [31] como técnica estadística de clasificación que discrimina de manera efectiva entre los distintos grupos de la variable a predecir. El objetivo es lograr la máxima separación transformando los datos a otro espacio compuesto por nuevas variables definidas como combinación de las variables originales. La determinación de la combinación lineal de variables se basa en la maximización de la dispersión entre los grupos y la minimización de la dispersión dentro de cada grupo (maximización de la función discriminante).

Su y Liu [108] formulan, a través del análisis discriminante, la mejor combinación lineal que maximiza el AUC bajo la suposición de normalidad multivariante, para las variables predictoras.

Teorema 1. - *Sean* $\mathbf{X_k}$ *los valores de p los biomarcadores para los individuos de la clase* $k = 0, 1$*, siguiendo una distribución normal:* $\mathbf{X_0} \sim N(\mu_0, \Sigma_0)$ *y* $\mathbf{X_1} \sim N(\mu_1, \Sigma_1)$*. La combinación lineal óptima cumple que:*

$$(\beta_0, \beta_1, \ldots, \beta_p) \propto (\Sigma_0 + \Sigma_1)^{-1} (\mu_1 - \mu_0) \tag{4.1}$$

Consecuentemente, se formula el AUC óptimo bajo el modelo propuesto por Su y Liu, como se enuncia en el siguiente corolario.

Corolario 1. *Dadas las condiciones del teorema 1, el AUC máximo viene dado por la siguiente expresión:*

$$AUC = \Phi(\sqrt{\mu^T (\Sigma_1 + \Sigma_0)^{-1} \mu}) \tag{4.2}$$

donde $\mu = \mu_1 - \mu_0$ y Φ es la función de distribución de la normal estándar.

4.1.2. Estimación no paramétrica

Si bien Su y Liu [108] presentan un modelo lineal que brinda la máxima capacidad de discriminación entre las dos clases (asociado a un AUC máximo), está condicionado al supuesto de normalidad multivariante. Esta suposición de normalidad a menudo no es fácil de cumplir en la práctica clínica real, ya que resulta ser demasiado exigente, en parte debido a la simetría que deben tener los biomarcadores. Sin embargo, para muchas enfermedades, las variables tienden a adoptar distribuciones asimétricas, donde

la progresión o etapas más avanzadas de la enfermedad están asociadas con valores elevados de las pruebas diagnósticas [19].

Esta limitación de cumplir con la suposición de normalidad resalta la necesidad de adoptar enfoques más flexibles que no estén condicionados por asunciones específicas sobre la distribución. Pepe y Thompson [83] abordaron esta limitación formulando un enfoque no paramétrico para estimar el modelo lineal que maximiza el AUC basado en la estadística U de Mann-Whitney [64]. Esta formulación y sus sugerencias han dado lugar al desarrollo de enfoques no paramétricos y semiparamétricos en la construcción de clasificadores bajo criterios de optimalidad derivados de la curva ROC. Específicamente, han sido la base para la formulación de los algoritmos presentados en esta tesis [4, 8].

En concreto, Pepe y Thompson [83] proponen el siguiente modelo lineal:

$$L_\beta(\mathbf{X}) = X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{4.3}$$

donde p denota el número de biomarcadores, X_i el biomarcador $i \in [1, \ldots, p]$ y β_i el parámetro a ser estimado. Considerando el vector de parámetros óptimos $\hat{\beta}$, se obtiene el máximo AUC empírico estimado con el estadístico U de Mann-Whitney, dado por la siguiente expresión, equivalente a la expresión (2.10):

$$\widehat{AUC} = \frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_0} I(L_{\hat{\beta}}(\mathbf{X_{1i}}) > L_{\hat{\beta}}(\mathbf{X_{0j}})) + \frac{1}{2} I(L_{\hat{\beta}}(\mathbf{X_{1i}}) = L_{\hat{\beta}}(\mathbf{X_{0j}}))}{n_0 \cdot n_1} \tag{4.4}$$

donde $\mathbf{X_{ki}}$ es el vector de variables para el individuo i del grupo $k = 0, 1$.

Obsérvese que el modelo lineal propuesto (4.3) no incluye ni el intercepto ni el coeficiente asociado a la variable X_1, como se presenta comúnmente en la expresión general del modelo lineal ($\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$). Esto resulta en la reducción de la estimación de $p-1$ parámetros. La justificación de esto radica en la propiedad de invariancia de la curva ROC frente a cualquier transformación monótona, que implica que el AUC de $L_\beta(\mathbf{X})$ (4.3) es equivalente al AUC de la expresión general de $p+1$ parámetros, tal como afirma el siguiente teorema.

Teorema 2. *Siendo X_1, X_2 variables predictoras, encontrar la combinación lineal $\beta_0 + \beta_1 X_1 + \beta_2 X_2$ con $\beta_1, \beta_2 \in \mathbb{R}$ que maximiza el AUC es equivalente a encontrar la combinación $X_1 + \beta X_2$ cuyo AUC es máximo.*

Demostración. La demostración del teorema es inmediata aplicando la propiedad de invarianza de la curva ROC (proposición 3). Las curvas ROC para $\beta_0 + \beta_1 X_1 + \beta_2 X_2$ y $X_1 + \beta X_2$ son equivalentes por ser resultado de aplicar transformaciones monótonas al dividir la primera expresión por β_1 y restar la constante β_0. □

Nótese que, al ser (4.4) una función no diferenciable, la estimación del modelo óptimo (4.3) implica explorar todo el espacio del vector de parámetros $\mathbb{R}^{p-1}$ y las posibles combinaciones de coeficientes y variables, lo cual resulta computacionalmente intratable. Pepe y Thompson [83] proponen una optimización discreta de los valores β_i en el intervalo $[-1, 1]$. Esto se debe a que el AUC de $X_i + \beta X_j$ para $\beta > 1$ y $\beta < -1$ es el mismo que el de $\alpha X_i + X_j$ para $\alpha = \frac{1}{\beta}$ en $[-1, 1]$, cubriendo todos los posibles valores en $\mathbb{R}$. En concreto, los autores sugieren una búsqueda en cuadrícula sobre 201 valores equidistantes en el intervalo $[-1, 1]$.

Aunque restringir la optimización al intervalo $[-1, 1]$ hace que la formulación inicial pueda ser computacionalmente viable, esta optimización todavía conlleva un coste computacional elevado para dimensiones $p \geq 3$. Específicamente, se trata de un problema NP-duro del orden de pk^{p-1}, donde p es el número de variables predictoras y k es la cantidad de valores posibles para cada coeficiente β_i. Para abordar esta limitación computacional, Pepe et al. [83, 82] sugirieron el uso de algoritmos con optimización paso a paso, seleccionando y estimando en cada paso la mejor combinación lineal de dos biomarcadores, e incorporando un nuevo biomarcador en cada iteración. De esta manera, el problema se vuelve computacionalmente abordable, ya que en cada paso se estima un solo coeficiente del modelo.

Esta estrategia de optimización parcial en cada paso fue implementada por Esteban et al. [29] y posteriormente por Kang et al. [48, 47] para la optimización del AUC. Aunque ambas propuestas siguen un enfoque de algoritmo paso a paso, siguiendo las propuestas de Pepe et al. [83, 82], difieren parcialmente en su enfoque. Esteban et al. proponen un método que busca en cada paso la combinación óptima de dos variables (par de biomarcadores y el coeficiente β), considerando el tratamiento de empates. Por otro lado, Kang et al. presentan un enfoque más simple y menos exigente, donde se establece un orden de inclusión de variables al inicio del algoritmo, basado en la ordenación de valores de AUC. El enfoque presentado por Kang et al. podría considerarse como un caso particular más sencillo del algoritmo paso a paso de Esteban et al., donde los nuevos biomarcadores que se incluyen en cada etapa están fijados desde el principio y donde no se consideran los empates.

Kang et al. [48, 47] proponen dos tipos de métodos paso a paso (métodos de paso ascendente y de paso descendente), aunque recomendaron el uso del procedimiento descendente. Sin pérdida de generalidad, a continuación se detalla el algoritmo paso a paso descendente:

1. Calcular la estimación empírica del AUC basado en la estadística U de Mann-Whitney (2.10) para cada uno de los p biomarcadores originales.

2. Ordenar estos p biomarcadores según los valores de las estimaciones empíricas del AUC de mayor a menor. Supongamos que el orden es el siguiente: $X_1, X_2, \ldots, X_p$.

3. Combinar los dos primeros biomarcadores (aquellos con los dos AUC más grandes) utilizando la búsqueda empírica de coeficientes propuesta por Pepe et al. [83], seleccionando el parámetro $\hat{\beta}$ óptimo de entre los 201 en el intervalo $[-1, 1]$ que maximiza el AUC:

$$\widehat{AUC} = \frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_0} I(X_{1i,1} + \hat{\beta}X_{1i,2} > X_{0j,1} + \hat{\beta}X_{0j,2}) + \frac{1}{2}I(X_{1i,1} + \hat{\beta}X_{1i,2} = X_{0j,1} + \hat{\beta}X_{0j,2})}{n_0 \cdot n_1}$$

   donde $X_{ki,1}, X_{ki,2}$ corresponde al valor de la primera y segunda variable, respectivamente, para el individuo i del grupo $k = 0, 1$.

4. Una vez elegido el parámetro que maximiza el AUC (supongamos la combinación lineal óptima $X_1 + \beta_2 X_2$), considerar dicha combinación lineal como una sola variable y combinarla con la tercera variable siguiendo el procedimiento del paso anterior. Es decir, encontrar el valor $\hat{\beta}$ óptimo de la combinación lineal $\hat{\beta}(X_1 + \beta_2 X_2) + X_3$ o $(X_1 + \beta_2 X_2) + \hat{\beta}X_3$.

5. Repetir el paso 4 para el resto de biomarcadores ($p - 3$ veces) hasta que todos ellos estén incluidos en el modelo.

El enfoque propuesto por Esteban et al. [29] difiere del de Kang et al. [48, 47] al comenzar en el paso 3, seleccionando la combinación lineal de dos variables predictoras considerando las p iniciales: $X_i + \hat{\beta}X_j, i \neq j = 1, \ldots, p$. Este proceso implica la búsqueda no solo del parámetro β óptimo, sino también del par de biomarcadores que produzca la combinación lineal parcial óptima (mayor AUC). Es evidente que, en cada paso, el máximo AUC puede obtenerse para más de una combinación lineal óptima, lo que da lugar a empates. Estos empates se consideran en cada etapa y se mantienen hasta que se resuelven en los pasos siguientes (si es posible) o hasta que finaliza el algoritmo. Este tratamiento de empates y mayor flexibilidad en la inclusión de las variables en cada paso, a pesar de requerir una búsqueda más extensa y exigente en términos computacionales, se espera que otorgue al algoritmo una mayor robustez en términos de su rendimiento. Aunque el tratamiento de empates puede ser más exigente en términos computacionales, la complejidad computacional del algoritmo es significativamente menor en comparación con el propuesto inicialmente por Pepe et al. [83, 82], reduciendo un problema de complejidad computacional exponencial pk^{p-1} a uno de tipo polinómico $k(p-1)(\frac{3}{2}p - 1)$.

4.1.3. Enfoque Min-Max

Liu et al. [56] también abordan la limitación computacional que exige el modelo lineal propuesto por Pepe et al. [83, 82]. Los autores proponen un enfoque no paramétrico denominado min-max, que tiene la ventaja de ser computacionalmente viable independientemente del número de biomarcadores originales. Este método se basa en combinar linealmente los valores mínimo y máximo de los p biomarcadores, involucrando la búsqueda de un solo coeficiente que maximice el AUC. En concreto, el objetivo es estimar el parámetro β de manera que la combinación

$$X_{\text{máx}} + \beta X_{\text{mín}} \tag{4.5}$$

maximice el AUC basado en la estadística de U de Mann-Whitney [64], donde $X_{\text{mín}}$ y $X_{\text{máx}}$ son los vectores de los valores mínimo y máximo de los p biomarcadores originales para cada individuo, respectivamente. O, específicamente, estimar el coeficiente β que maximice la siguiente expresión:

$$\widehat{AUC} = \frac{\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} I(X_{1j,\max} + \beta X_{1j,\min} > X_{0i,\max} + \beta X_{0i,\min}) + \frac{1}{2} I(X_{1j,\max} + \beta X_{1j,\min} = X_{0i,\max} + \beta X_{0i,\min})}{n_0 \cdot n_1} \tag{4.6}$$

donde $X_{ki,\max} = \text{máx}_{1\leq j\leq p} X_{kij}$ y $X_{ki,\min} = \text{mín}_{1\leq j\leq p} X_{kij}$ para $k = 0, 1$ y cada individuo $i = 1, \ldots, n_k$.

Los autores proponen este enfoque basado en los siguientes razonamientos. Por definición, para todo biomarcador $j \in [1, p]$, siendo c el punto de corte, la sensibilidad satisface que

$$P(X_{1i,\min} > c) \leq P(X_{1ij} > c) \leq P(X_{1i,\max} > c) \tag{4.7}$$

para todos individuos $i = 1, \ldots, n_1$. De manera análoga, la especificidad satisface que

$$P(X_{0i,\max} \leq c) \leq P(X_{0ij} \leq c) \leq P(X_{0i,\min} \leq c) \tag{4.8}$$

Estos resultados indican que el biomarcador máximo $X_{\text{máx}}$ tiene una sensibilidad mayor y una especificidad menor para cualquier punto de corte c en comparación con cualquier biomarcador original j; y de forma inversa, lo cumple el biomarcador mínimo $X_{\text{mín}}$. Por tanto, si el objetivo es maximizar la sensibilidad (especificidad) de forma independiente, el biomarcador máximo (mínimo) es plausible. Sin embargo, esto no suele ocurrir en la práctica clínica. Lo que suele tener interés es maximizar un compromiso entre ambas. En este sentido, los autores se centran en la combinación de estos biomarcadores máximo y mínimo, lo que resulta en el enfoque que proponen. Para utilizar este algoritmo correctamente, los autores advierten que si los biomarcadores no se miden en las mismas unidades, es necesario normalizar los valores de cada biomarcador antes de aplicar el enfoque.

En conclusión, Liu et al. [56] proponen el enfoque de combinación min-max que es de naturaleza no paramétrica, siendo más robusto que los modelos paramétricos, como así lo demuestran los autores. Este enfoque no paramétrico reduce el problema a estimar un solo coeficiente sea cual sea el número de biomarcadores originales, siendo siempre eficiente computacionalmente. La estimación del paramétro β óptimo puede realizarse utilizando diferentes métodos. Por ejemplo, se pueden seguir las recomendaciones de Pepe et al. [83, 82], y buscar el valor óptimo de β entre los 201 valores en el rango de $[-1, 1]$, de forma que se maximice el AUC estimado (4.6).

4.2. Modelos con criterio de optimalidad derivados de la curva ROC

Aunque la combinación de múltiples biomarcadores puede mejorar la precisión diagnóstica, las estrategias varían según el objetivo, ya sea maximizar el área o el área parcial bajo la curva ROC (pAUC), el índice de Youden, o buscar maximizar la sensibilidad en un rango específico de especificidades, entre otros. Es importante tener en cuenta que una solución óptima para un objetivo puede no ser adecuada para otro. En la sección anterior se han presentado algoritmos centrados en optimizar el AUC, que han sido la base teórica o punto de partida para el desarrollo de algunos algoritmos posteriores con criterio de optimalidad derivados de la curva ROC. Específicamente, en esta sección se presentan estudios desarrollados para la combinación lineal de biomarcadores, enfocados en la optimización de otros parámetros adicionales derivados de la curva ROC.

Liu et al. [55] investigaron la combinación lineal óptima de marcadores diagnósticos para maximizar la sensibilidad dentro de un rango específico de especificidades. Su estudio se basó en el trabajo previo de Su y Liu [108], identificando posibles limitaciones en el enfoque, puesto que estas combinaciones lineales podrían mostrar una sensibilidad inadecuada en ciertos rangos de especificidad deseados. Centrándose en aplicaciones donde no se opera en todo el rango de la curva, sino solo en regiones específicas, otros autores han comparado los enfoques existentes e investigado y desarrollado otros métodos. Por ejemplo, autores como Hsu y Hsueh [44] o Yu y Park [126] propusieron métodos para maximizar el pAUC paramétrico. Sin embargo, estos enfoques están limitados por la suposición de normalidad. Posteriormente, Yan et al. [121] propusieron enfoques tanto paramétricos como no paramétricos para la combinación lineal de múltiples biomarcadores con el fin de maximizar el pAUC. Compararon el rendimiento de sus propuestas y otros métodos, que incluyeron el enfoque de Liu et al. [55], así como otros métodos presentados como el enfoque de Su y Liu [108], y métodos de paso a paso

como el de Kang et al. [48] Más recientemente, Ma et al. [63] propusieron un método no paramétrico que extiende el método min-max [56] para la estimación del pAUC en biomarcadores continuos. También compararon su rendimiento con los enfoques de Su y Liu [108], Li et al. [55] y la regresión logística.

Aunque en el contexto clínico de problemas de clasificación binaria, los investigadores han prestado mayor atención a evaluar la precisión diagnóstica mediante el AUC, el índice de Youden también se ha utilizado en diversos estudios clínicos. Esto ocurre especialmente cuando no hay un consenso claro sobre si se debe dar prioridad a la sensibilidad o a la especificidad. En tales casos, el índice de Youden ofrece una métrica adecuada como resumen del diagnóstico, además de servir como criterio para seleccionar el punto de corte óptimo que dicotomiza el biomarcador [84], lo cual es fundamental para el diagnóstico de la enfermedad. Yin y Tian [122] abordaron su investigación con la premisa de considerar el índice de Youden, además del AUC, debido a su capacidad de ofrecer una medida directa y significativa de la precisión diagnóstica en el punto de corte óptimo. Como un primer paso hacia este objetivo, exploraron la optimización simultánea del AUC y el índice de Youden, presentando tanto enfoques paramétricos como no paramétricos para estimar la región de confianza conjunta de ambas medidas. Además, en un estudio adicional, Yin y Tian [123] adoptaron el índice de Youden como criterio para determinar el punto de corte de diagnóstico, proponiendo enfoques paramétricos y no paramétricos para estimar la región de confianza conjunta de sensibilidad y especificidad en dicho punto de corte.

Considerando el índice de Youden como función objetivo para buscar la combinación lineal óptima, Yin y Tian [124] llevaron a cabo un estudio comparativo de diversos enfoques existentes. Los autores proponen un enfoque de combinación paso a paso, basado en el método de paso a paso descendente de Kang et al. [48], presentado en la sección anterior. Lo compararon con otros métodos como el enfoque min-max [56], el enfoque paramétrico del índice de Youden (2.30-2.33) y el enfoque no paramétrico de suavizado tipo Kernel (2.37). Además, llevaron a cabo una comparación empírica de las tasas de clasificación general óptimas entre la combinación propuesta basada en el índice de Youden y la tradicional basada en el AUC. Los autores demostraron una ganancia significativa en la precisión diagnóstica para la combinación propuesta.

Con el objetivo de abordar las limitaciones de los métodos existentes, llevamos a cabo los estudios presentados en este compendio [4, 8, 7]. Nos basamos en la eficacia del índice de Youden como métrica diagnóstica y criterio de selección del punto de corte óptimo, así como en su menor atención recibida en la literatura en comparación con el AUC. En estos artículos, presentamos enfoques diseñados para maximizar el índice de Youden (enfoque paso a paso, enfoque min-max-median/IQR), utilizando como base

teórica los estudios previos desarrollados para optimizar el AUC, presentados en la sección anterior. Los algoritmos propuestos se presentan en detalle en los capítulos siguientes (capítulos 5-6).

4.3. Otros algoritmos

En esta sección se introducen otros algoritmos adicionales para la clasificación binaria de enfermedades, cuyos criterios a optimizar son diferentes a las métricas derivadas de la curva ROC. No obstante, estos algoritmos pueden ser evaluados posteriormente utilizando el índice de Youden u otras métricas comunes, lo que permite realizar estudios comparativos entre todos los enfoques. Esto permite ampliar el abanico de posibilidades, incluyendo clasificadores lineales y no lineales, con el fin de seleccionar el algoritmo más adecuado según los datos y el problema en cuestión. A continuación, se presentan los algoritmos que se han utilizado en el trabajo de la tesis, ya sea en la comparación con los algoritmos propuestos [4, 8, 7] o para abordar diferentes problemas de predicción en el ámbito de la salud [97, 5, 34].

4.3.1. Modelos lineales

Los modelos de regresión lineales, que estiman relaciones entre la variable dependiente y las independientes o sus transformaciones, han sido ampliamente utilizados en la práctica clínica debido a varias razones. En primer lugar, ofrecen una fácil interpretación de la relación entre las variables, lo que es crucial en entornos médicos donde se necesita comprender claramente el impacto de cada variable. Además, estos modelos suelen ser algoritmos simples de implementar y computacionalmente eficientes, lo que los hace accesibles y adecuados para el análisis de grandes conjuntos de datos médicos. Aunque no pueden capturar relaciones no lineales o complejas entre las variables, su robustez ha sido demostrada en diversas situaciones prácticas.

En el contexto de problemas de clasificación, surge la conocida regresión logística [117], que modela la probabilidad de un evento (enfermedad o no enfermedad) dado un conjunto de variables independientes $(X_1, \ldots, X_p)$ a través de la función logística. Por tanto, el problema de clasificación se convierte en la estimación de los parámetros $(\beta_0, \beta_1, \ldots, \beta_p)$ tal que:

$$P(\mathbf{Y} = 1|\mathbf{X} = x) = \frac{1}{1 + e^{-\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} \tag{4.9}$$

de forma que su dependencia lineal viene dada por la siguiente expresión:

$$log\frac{P(\mathbf{Y} = 1|\mathbf{X} = x)}{1 - P(\mathbf{Y} = 1|\mathbf{X} = x)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{4.10}$$

La estimación de los parámetros $(\beta_0, \beta_1, \ldots, \beta_p)$ se realiza generalmente mediante el método de máxima verosimilitud, el cual busca maximizar la verosimilitud de los datos observados respecto a los parámetros del modelo. A pesar de que la función objetivo de la regresión logística es la función de máxima verosimilitud logística en lugar del índice de Youden, la hemos incluido en nuestros estudios [4, 8, 7] como parte de la comparación de algoritmos, debido a que es la técnica más comúnmente utilizada para combinar linealmente múltiples biomarcadores. También se ha utilizado en el trabajo [34], presentado en la sección 7.3, para construir combinaciones lineales genéticas y evaluar su asociación con el riesgo de adenomas colorrectales.

Los modelos de regresión lineal, además de asumir una relación lineal presuponen independencia en los errores de las observaciones. Sin embargo, en la práctica, los datos pueden mostrar estructuras más complejas, con observaciones agrupadas en diferentes niveles o unidades, y la variabilidad puede diferir entre estos grupos. Este escenario es común en estudios longitudinales, donde se realiza un seguimiento continuo de un grupo de individuos a lo largo del tiempo, recopilando información sobre su estado de salud, características u otros aspectos. En tales escenarios, los datos de cada individuo a lo largo del tiempo suelen estar correlacionados, violando la suposición de independencia entre las observaciones. Para abordar estas situaciones, se desarrollaron los modelos lineales mixtos [114], que permiten capturar la estructura de correlación y heterogeneidad presente en los datos, mediante la inclusión de una combinación de efectos fijos y aleatorios como variables predictoras. En el diseño de modelos lineales mixtos es crucial distinguir entre los efectos fijos y los efectos aleatorios. Mientras que los efectos fijos representan relaciones constantes aplicables a todas las unidades de observación, los efectos aleatorios varían y se utilizan para capturar la variabilidad entre los grupos o niveles. Esta distinción puede resultar difícil y generar controversias [53], pero es fundamental para modelar adecuadamente la variabilidad y las correlaciones presentes en los datos longitudinales. En esencia, los efectos fijos representan relaciones promedio entre variables, mientras que los efectos aleatorios capturan la variabilidad no explicada por los efectos fijos y permiten modelar la heterogeneidad entre grupos o niveles de agrupamiento en los datos. La fórmula general de un modelo lineal mixto, expresada en forma matricial, se presenta como:

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\beta + \mathbf{Z}b + \epsilon \\ \epsilon &\sim \mathcal{N}(0, \Sigma) \\ b &\sim \mathcal{N}(0, D) \end{aligned} \tag{4.11}$$

donde $\mathbf{Y}$ es el vector de respuestas observadas, $\mathbf{X}$ la matriz de diseño para los efectos fijos, β el vector de coeficientes para los efectos fijos, $\mathbf{Z}$ la matriz de diseño para los

efectos aleatorios, b el vector de coeficientes de efectos aleatorios y ϵ el vector de errores.

Un ejemplo ilustrativo de la utilidad de los modelos lineales mixtos se encuentra en el estudio presentado en el artículo [97], incluido en este compendio de tesis y detallado en la sección 7.2.

4.3.2. Técnicas de aprendizaje automático

En los últimos años, los algoritmos de aprendizaje automático (Machine Learning, ML en inglés) y aprendizaje profundo (Deep Learning, DL en inglés) han sido aplicados en diversos campos [94], incluyendo la práctica clínica y la investigación médica, gracias a su capacidad para ofrecer alto rendimiento y eficiencia, así como por su habilidad para capturar relaciones no lineales [103, 78, 74]. En el ámbito biomédico, estos algoritmos han sido empleados para predecir eventos, diagnosticar enfermedades y pronosticar cánceres, entre otros usos [102], lo que ha resultado en herramientas útiles para la toma de decisiones. Entre los algoritmos más comúnmente utilizados se encuentran los árboles de clasificación, los modelos ensemble como el RandomForest y el XGBoost (eXtreme Gradient Boosting), las máquinas de vector suporte (Support Vector Machine, SVM en inglés) y las redes neuronales como los perceptrones multicapa (MLP, siglas en inglés). Especialmente el algoritmo XGBoost, desarrollado por Chen y Guestrin [21], destaca como uno de los algoritmos de aprendizaje automático más populares y efectivos en los últimos años. Ha demostrado buenos resultados, siendo líder en algunos estudios de vanguardia [11]. Estos modelos de ML requieren la configuración de parámetros adicionales conocidos como hiperparámetros, los cuales definen su arquitectura. Estos hiperparámetros deben ser establecidos antes del entrenamiento del modelo y pueden ajustarse utilizando técnicas de optimización de búsqueda de hiperparámetros para maximizar el rendimiento del modelo.

Aunque tanto el Random Forest como el XGBoost son algoritmos que se basan en la construcción de múltiples árboles de clasificación, difieren en su enfoque. En lugar de construir varios árboles de forma independiente como el Random Forest (bagging), el XGBoost construye los árboles secuencialmente (boosting), centrándose en corregir los errores del modelo en cada paso. Específicamente, el XGBoost es una implementación optimizada del aumento de gradiente, basada en el principio de aprendizaje de conjunto en orden secuencial, donde los errores se minimizan (función de pérdida) utilizando un algoritmo de descenso de gradiente. La idea subyacente es entrenar varios modelos base débiles secuencialmente para crear un clasificador fuerte con mayor precisión. Considerando un conjunto de modelos de árboles de decisión f, la función de pérdida

a minimizar en la iteración t se expresa de la siguiente manera:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y_i}^{(t-1)} + f_t(\mathbf{X}_i)\right) + \Omega(f_t) \tag{4.12}$$

donde l es el término de pérdida, y_i representa la salida real, $\hat{y_i}^{(t-1)}$ es la predicción del i-ésimo individuo en la $(t-1)$-ésima iteración del modelo, y $\Omega(f) = \gamma T + \frac{1}{2}\lambda||\omega||^2$ es el término de regularización, que penaliza la complejidad del modelo para evitar el sobreajuste y depende de valores de hiperparámetros.

A pesar de ser uno de los algoritmos de aprendizaje automático más ampliamente utilizados por su buen rendimiento, el XGBoost no siempre ha superado a los métodos estadísticos convencionales, como la regresión logística [50, 24, 116], por ejemplo. Esto evidencia el hecho de que, aunque un algoritmo pueda mostrar resultados prometedores en el estado del arte, la selección del modelo óptimo depende del problema y las características de los datos. Por lo tanto, es recomendable realizar un estudio comparativo de diversas técnicas para analizar su rendimiento y seleccionar la más adecuada para el problema específico en cuestión. En los artículos [5, 7] de este compendio comparamos los resultados de diferentes algoritmos con el fin de seleccionar el óptimo para el objetivo del estudio. En concreto, en [5] (sección 7.1) exploramos la capacidad de los algoritmos ensemble (Random Forest y XGBoost) y MLP para la predicción de severidad en pacientes con COVID-19. En [7] (sección 6.2), evaluamos el rendimiento de nuestros enfoques propuestos en comparación con el algoritmo XGBoost, ampliando el análisis más allá de las técnicas tradicionales.

Hasta aquí se ha abordado la teoría que respalda el desarrollo de los estudios publicados. A continuación, se presentan los siguientes capítulos enfocados en los propios artículos que conforman esta tesis. Cada capítulo proporciona una descripción y resultados generales del trabajo, además de incluir el artículo completo publicado.

Capítulo 5

Algoritmo paso a paso bajo maximización del índice de Youden

En este capítulo se presenta el primer artículo del compendio de esta tesis [4], cuya investigación ha sido también la base de otros desarrollos que han dado lugar a los artículos presentados en el siguiente capítulo. En concreto, nuestro estudio se centra en la propuesta de un algoritmo paso a paso no paramétrico para problemas de clasificación binaria con criterio de optimalidad derivado de la curva ROC. A continuación, se detallan las contribuciones de este trabajo y las razones que impulsaron el desarrollo de la propuesta.

Proponemos un enfoque que busca una estimación basada en la maximización del índice de Youden. La elección de esta métrica se justifica por diversas razones. En primer lugar, el índice de Youden es una métrica adecuada para medir la precisión del diagnóstico cuando no hay un consenso o razón clara para proporcionar valores más altos de sensibilidad o especificidad, proporcionando un equilibrio adecuado. Asimismo, el índice de Youden ofrece un buen criterio para elegir el mejor punto de corte para dicotomizar un biomarcador, lo que resulta importante en términos de utilidad clínica. Por otro lado, a diferencia de otras métricas como el AUC, el estudio y exploración de métodos que optimizan el índice de Youden, en nuestra opinión, no ha recibido suficiente atención en la literatura y, por tanto, es necesario analizar y desarrollar enfoques que optimicen esta métrica así como compararlos con otros del estado del arte.

Nuestro estudio parte de investigaciones previas de la literatura, con el objetivo de aprovechar sus ventajas y abordar sus limitaciones. En concreto, nos apoyamos en los estudios de Pepe et al. [83, 82], que proponen un enfoque libre de distribución para estimar un modelo lineal que maximiza el AUC. Su formulación reduce la estimación de parámetros en comparación con el modelo lineal comúnmente conocido, gracias a la propiedad de invarianza de la curva ROC. Para la estimación de los parámetros óptimos

del modelo, los autores sugieren una optimización discreta que implica una búsqueda en cuadrícula de 201 valores, reduciendo así la carga computacional. Sin embargo, este enfoque conlleva una carga computacional elevada cuando el número de biomarcadores aumenta. Para afrontar esta limitación, sugieren aplicar métodos de paso a paso.

Nuestro enfoque aborda la limitación computacional considerando esta sugerencia. El índice de Youden es una métrica que deriva también de la curva ROC y, por tanto, se pueden aplicar los resultados de Pepe et al. [83, 82] a nuestra propuesta. En concreto, proponemos un enfoque no paramétrico paso a paso que adapta el propuesto por Esteban et al. [29] para la maximización del índice de Youden, utilizando la formulación y búsqueda discreta de Pepe et al. La idea general es seleccionar la mejor combinación lineal de dos variables en cada iteración, incluyendo una nueva variable en cada paso. Nuestra propuesta no establece el orden de entrada de las variables y considera los empates que pueden darse en cada paso, puesto que el índice de Youden máximo puede alcanzarse para más de una combinación lineal óptima. Esto hace que nuestro enfoque sea robusto, considerando una optimización más precisa que otros enfoques paso a paso.

Aunque nuestro enfoque es computacionalmente más abordable al requerir la estimación de un solo parámetro en cada paso, esta estimación óptima en cada paso no garantiza la optimización global del índice de Youden. Por ello, se realizó un análisis del comportamiento del algoritmo en diferentes escenarios con el fin de conocer su adecuación. Además, se comparó con otras técnicas de la literatura para determinar la más adecuada según el escenario.

Nuestro enfoque fue comparado con una amplia gama de técnicas, incluyendo el enfoque paso a paso de Yin y Tian [124], el método min-max [56], el algoritmo clásico de regresión logística, y los enfoques paramétrico y no paramétrico de tipo Kernel del índice de Youden, presentados en la sección 2.2. El enfoque de Yin y Tian podría considerarse como un caso particular más simple de nuestro enfoque propuesto, donde los nuevos biomarcadores de cada etapa se fijan desde el principio y donde los empates no se consideran. El método min-max se adaptó para optimizar el índice de Youden, y el rendimiento de la regresión logística fue evaluado con el índice de Youden tras su estimación, por tener una función objetivo diferente en su proceso de estimación. Para la estimación de los parámetros de las combinaciones lineales de los enfoques paramétrico y no paramétrico de tipo Kernel, se utilizaron métodos de optimización numérica, por ser la expresión del índice de Youden una función continua y diferenciable (2.30)),(2.37).

Nuestro trabajo consideró la comparación de estos métodos en un amplio rango de escenarios simulados. Se consideraron diferentes distribuciones conjuntas y marginales

para los biomarcadores, desde distribuciones normales hasta no normales, con el objetivo de evaluar y comparar los métodos más allá de la normalidad. Se analizaron diferentes capacidades de discriminación entre los biomarcadores, así como distintas correlaciones y matriz de covarianzas para el grupo con enfermedad y sanos. Para cada escenario, consideramos diferentes tamaños muestrales con el fin de evaluar los métodos en tamaños de muestra pequeños y más grandes. El rendimiento de los métodos se analizó también en dos conjuntos de datos reales relacionados con casos de diagnóstico clínico (distrofia muscular de Duchenne y cáncer de próstata) a través de sus respectivos biomarcadores.

Los resultados mostraron que nuestro enfoque propuesto superó a los demás métodos en escenarios simulados con distribuciones marginales no normales, así como en el conjunto de datos de cáncer de próstata. Este conjunto de datos incluye variables, como el PSA, que muestran asimetrías claras que reflejan la progresión de la enfermedad. Todos los resultados obtenidos se presentan en el artículo, donde se discuten en detalle y se sugiere el uso de los diferentes métodos analizados según el escenario en cuestión.

En términos de tiempo computacional, los enfoques paso a paso y, en particular, nuestra propuesta, implican un tiempo computacional significativamente mayor al resto de algoritmos debido al tratamiento de empates, que pueden ser más comunes en tamaños de muestra pequeños. La carga computacional puede aumentar también a medida que el número de biomarcadores aumenta, debido al aumento de posibles combinaciones de variables. Esto puede ser una restricción en el uso de nuestro algoritmo. Otros algoritmos, como el enfoque min-max, tienen la ventaja de ser más eficientes. Sin embargo, los resultados del trabajo han demostrado que, generalmente, el enfoque min-max tiene una peor capacidad de discriminación.

En resumen, esta investigación ha permitido el desarrollo de un enfoque paso a paso que aborda las limitaciones de otras propuestas aparecidas en la literatura, centrado en la utilidad clínica y considerando las propiedades de la curva ROC bajo la maximización del índice de Youden. Hemos analizado su comportamiento en un amplio rango de escenarios y con diversos métodos de la literatura, lo que proporciona información útil y ciertas pautas para la selección del método más adecuado.

A continuación se presenta el artículo publicado, donde se detalla el trabajo en mayor profundidad.

mathematics

MDPI

Article

A Stepwise Algorithm for Linearly Combining Biomarkers under Youden Index Maximization

Rocío Aznar-Gimeno [1,*], Luis M. Esteban [2,*], Rafael del-Hoyo-Alonso [1], Ángel Borque-Fernando [3] and Gerardo Sanz [4]

[1] Department of Big Data and Cognitive Systems, Instituto Tecnológico de Aragón (ITAINNOVA), 50018 Zaragoza, Spain; rdelhoyo@itainnova.es
[2] Department of Applied Mathematics, Escuela Universitaria Politécnica de La Almunia, Universidad de Zaragoza, La Almunia de Doña Godina, 50100 Zaragoza, Spain
[3] Department of Urology, Hospital Universitario Miguel Servet and IIS-Aragón, Paseo Isabel La Católica 1-3, 50009 Zaragoza, Spain; aborque@comz.org
[4] Department of Statistical Methods and Institute for Biocomputation and Physics of Complex Systems-BIFI, University of Zaragoza, 50009 Zaragoza, Spain; gerardo.sanz@unizar.es
* Correspondence: raznar@itainnova.es (R.A.-G.); lmeste@unizar.es (L.M.E.)

Abstract: Combining multiple biomarkers to provide predictive models with a greater discriminatory ability is a discipline that has received attention in recent years. Choosing the probability threshold that corresponds to the highest combined marker accuracy is key in disease diagnosis. The Youden index is a statistical metric that provides an appropriate synthetic index for diagnostic accuracy and a good criterion for choosing a cut-off point to dichotomize a biomarker. In this study, we present a new stepwise algorithm for linearly combining continuous biomarkers to maximize the Youden index. To investigate the performance of our algorithm, we analyzed a wide range of simulated scenarios and compared its performance with that of five other linear combination methods in the literature (a stepwise approach introduced by Yin and Tian, the min-max approach, logistic regression, a parametric approach under multivariate normality and a non-parametric kernel smoothing approach). The obtained results show that our proposed stepwise approach showed similar results to other algorithms in normal simulated scenarios and outperforms all other algorithms in non-normal simulated scenarios. In scenarios of biomarkers with the same means and a different covariance matrix for the diseased and non-diseased population, the min-max approach outperforms the rest. The methods were also applied on two real datasets (to discriminate Duchenne muscular dystrophy and prostate cancer), whose results also showed a higher predictive ability in our algorithm in the prostate cancer database.

Keywords: linear combination; stepwise algorithm; Youden index; biomarkers; diagnosis

MSC: 62H30; 62J12; 62P10

Citation: Aznar-Gimeno, R.; Esteban, L.M.; del-Hoyo-Alonso, R.; Borque-Fernando, Á.; Sanz, G. A Stepwise Algorithm for Linearly Combining Biomarkers under Youden Index Maximisation. *Mathematics* **2022**, *10*, 1221. https://doi.org/10.3390/math10081221

Academic Editor: Jiancang Zhuang

Received: 22 February 2022
Accepted: 5 April 2022
Published: 8 April 2022

1. Introduction

In clinical practice, it is usual to obtain information on multiple biomarkers to diagnose diseases. Combining them into a single biomarker is a widespread practice that often provides better diagnostics than each of the biomarkers alone [1–6]. Although recent studies have analyzed the diagnostic accuracy of built models in the presence of covariates and binary biomarkers [7–9], the combination of continuous biomarkers should provide a better discrimination ability. Linear combination methods have been widely developed and applied for both binary and multi-class classification problems in medicine [10,11] for their ease of interpretation and good performance. The accuracy of a diagnostic marker is usually analyzed using statistics derived from the receiver operating characteristic (ROC) curve, such as sensitivity and specificity, the area or partial area under the ROC curve or the Youden index, which allow for its discriminatory capacity to be measured.

The formulation of algorithms to estimate binary classification models that maximize the area under the ROC curve is a widely developed line of research. Su and Liu [12], using a discriminant function analysis, provided the best linear combination that maximizes area under ROC curve (AUC) under the multivariate normality assumption. Pepe and Thompson [13] proposed a distribution-free approach to estimate the linear model that maximizes AUC based on the Mann–Whitney U statistic [14]. This formulation has given rise to the development of non-parametric and semiparametric approaches in the construction of classifiers under optimality criteria derived from the ROC curve.

The process that Pepe and Thompson proposed lies in a discrete optimization that is based on a grid search over the parameter vector for a set of selected values. However, this process requires a great computational effort when the number of biomarkers is greater than or equal to three. In order to address the computational burden, various methods were proposed. Liu et al. [15] developed a non-parametric min-max approach, reducing the problem to a linear combination of two markers (minimum and maximum of biomarker values). Pepe et al. [13,16] also suggested the use of stepwise algorithms, where a new variable is introduced into the model at each stage searching for the partial combination of variables that maximizes AUC. Esteban et al. [17] implemented this approach, providing strategies to handle ties that appear in the sequencing of partial optimizations. Kang et al. [10,18] proposed a less computationally demanding stepwise approach based on a descending order of the AUC values corresponding to the predictor variables.

Other authors have developed algorithms focused on optimizing other parameters derived from the ROC curve. Liu et al. [19] analyzed the optimal linear of diagnostic markers that maximize sensitivity over a range of specificity. Yin and Tian [20] analyzed the joint inference on sensitivity and specificity at the optimal cut-off point associated with the Youden index. More recently, Yu and Park [21], Yan et al. [22] and Ma et al. [23] explored methods for the linear combination of multiple biomarkers that optimizes the pAUC.

The Youden index has also been used in different clinical studies and is both an appropriate summary for making the diagnosis and a good criterion for choosing the best cut-off point to dichotomize a biomarker [24]. The Youden index defines the effectiveness of a biomarker, as it maximize the sum of the sensitivity and specificity when an equal weight is given for both values [25]. Thus, the cut off point that simultaneously maximizes the probability of correctly classifying positive and negative subjects or minimizes the maximum of the misclassification error probability is chosen. It ranges from 0 to 1, where a 0 value indicates that the biomarker is equally distributed on the positive and the negative populations, whereas a value of 1 indicates completely separate distributions [26].

Based on the stepwise approach of Kang et al. [18], Yin and Tian [27] carried out a study aimed at optimizing the Youden index. These authors also analyzed the optimization of the AUC and Youden index simultaneously and presented both a parametric and a non-parametric approach to estimate the joint confidence region for the AUC and the Youden index [28]. However, the usual procedure is to estimate models that maximize either the AUC or the Youden index separately.

Unlike the AUC, the study and exploration of methods that optimize the Youden index has not received enough attention in the literature. The aim of our study was to propose a new stepwise distribution-free approach to find the optimal linear combination of continuous biomarkers based on maximizing the Youden index. In order to analyze its performance, our method was compared with five other linear methods from the literature (the Yin and Yan stepwise approach, the min-max method, logistic regression, a parametric approach under multivariate normality and a non-parametric kernel smoothing approach) adapted to optimize the Youden index, both in simulated data and in real datasets.

2. Materials and Methods

Firstly, we introduce the non-parametric formulation of Pepe et al. [13,16] and their suggestions for the estimation of the parameter vector of the linear model, which are the basis for the formulation and estimation of our proposed algorithm and of the analyzed algorithms. Then, we introduce our proposed method and five existing models in the

literature adapted to optimize the Youden index to be compared: stepwise algorithm proposed by Yin and Tian, min-max approach, logistic regression, parametric method under multivariate normality and non-parametric kernel smoothing method. Finally, the simulated scenarios, as well as the real datasets considered, are described. All methods were programmed and applied using free software R [29]. In particular, a library in R (***SLModels***) [30] openly available to the scientific community was created that incorporates our proposed stepwise algorithm, among other linear algorithms.

Suppose that p continuous biomakers are measured for n_1 individuals with disease: $\mathbf{X_1} = (\mathbf{X_{11}}, \ldots, \mathbf{X_{1n_1}})$ and for n_2 individuals without it: $\mathbf{X_2} = (\mathbf{X_{21}}, \ldots, \mathbf{X_{2n_2}})$. $\mathbf{X_{ki}}$ denotes the vector of p biomarkers for the i^{th} individual of group k = 1,2 (disease and non-disease) and X_{kij} the j^{th} biomarker ($j = 1, \ldots, p$) for the i^{th} individual of group k = 1,2.

Given $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ as the parameter vector, the linear combination for the disease and non-disease group is represented as follows:

$$\mathbf{Y_k} = \boldsymbol{\beta}^T \mathbf{X_k}, \;\; k = 1, 2 \tag{1}$$

The Youden index (J) is defined as

$$\begin{aligned} J &= max_c\{Sensitivity(c) + Specificity(c) - 1\} \\ &= max_c\{F_{\mathbf{Y_2}}(c) - F_{\mathbf{Y_1}}(c)\} \end{aligned} \tag{2}$$

where c denotes the cut-off point and $F_{\mathbf{Y_k}}(c) = P(\mathbf{Y_k} \leq c)$ the cumulative distribution function of random variable $\mathbf{Y_k}, k = 1, 2$.

Denoting by $c_{\boldsymbol{\beta}} = \{c : max_c\big(F_{\mathbf{Y_2}}(c) - F_{\mathbf{Y_1}}(c)\big)\}$ as the optimal cut-off point and substituting (1) in (2), the empirical estimate of Youden index ($\hat{J}_{\boldsymbol{\beta}}$) is obtained as follows:

$$\begin{aligned} \hat{J}_{\boldsymbol{\beta}} &= \hat{F}_{\mathbf{Y_2}}(\hat{c}_{\boldsymbol{\beta}}) - \hat{F}_{\mathbf{Y_1}}(\hat{c}_{\boldsymbol{\beta}}) \\ &= \frac{\sum_{i=1}^{n_2} I\big(\boldsymbol{\beta}^T \mathbf{X_{2i}} \leq \hat{c}_{\boldsymbol{\beta}}\big)}{n_2} - \frac{\sum_{i=1}^{n_1} I\big(\boldsymbol{\beta}^T \mathbf{X_{1i}} \leq \hat{c}_{\boldsymbol{\beta}}\big)}{n_1} \end{aligned} \tag{3}$$

where I denotes the indicator function.

2.1. Background: Non Parametric Approach

By contrast to Su and Liu [12], who provided best linear model under multivariate normality, Pepe and Thompson [13] proposed a non-parametric approach to estimate the linear model that maximizes the AUC evaluated by the Mann–Whitney U statistic,

$$\widehat{AUC} = \frac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I(L(\mathbf{X_{1i}}) > L(\mathbf{X_{2j}})) + \frac{1}{2} I(L(\mathbf{X_{1i}}) = L(\mathbf{X_{2j}}))}{n_1 \cdot n_2} \tag{4}$$

considering the linear model formulation as follows:

$$L(\mathbf{X}) = X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{5}$$

where p denotes the number of biomarkers, X_i the biomarker $i \in [1, \ldots, p]$ and β_i the parameter to be estimated. In order to be able to address the computational burden, Pepe et al. [13,16] suggest, for the estimation of the parameter β_i, a discrete optimization that is based on a grid search over 201 equally spaced values in the interval $[-1,1]$. The justification for choosing this range lies in the property of the ROC curve that is invariant to any monotonic transformation. Consider, for simplicity, the linear combination of biomarkers $X_i + \beta X_j$. Then, due to the invariant property of the ROC curve for any monotonic transformation, dividing by the β value does not change the value of the sensitivity and specificity pair. That is, estimating $X_i + \beta X_j$ for $\beta > 1$ and $\beta < -1$ is equivalent to estimating $\alpha X_i + X_j$ for $\alpha = \frac{1}{\beta} \in [-1, 1]$ and, therefore, all possible values of $\beta \in \mathbb{R}$ are covered.

However, the search for the best linear combination in (5) is still computationally costly when $p \geq 3$. To solve this problem, Pepe et al. [13,16] suggested using stepwise algorithms by turning a computationally intractable problem into an approachable problem of single-parameter estimation (linear combination of two variables) $p-1$ times.

2.2. Our Proposed Stepwise Approach

Our proposed stepwise linear modelling (SLM) is an adaptation of the one proposed by Esteban et al. [17] for Youden index maximization. The general idea of this approach, as Pepe et al. [13,16] suggest, is to follow a step by step algorithm that includes a new variable in each step, selecting the best combination (or combinations) of two variables, in terms of maximizing the Youden index. The following steps explain the algorithm in detail:

1. Firstly, given p biomarkers, the linear combination of the two biomarkers that maximizes the Youden index is chosen,

$$\hat{J}_{\beta_2} = \frac{\sum_{i=1}^{n_2} I\left(X_{2ij} + \beta_2 X_{2ik} \leq \hat{c}_{\beta_2}\right)}{n_2} - \frac{\sum_{i=1}^{n_1} I\left(X_{1ij} + \beta_2 X_{1ik} \leq \hat{c}_{\beta_2}\right)}{n_1} \qquad \beta_2 \in [-1,1], \quad \forall j \neq k = 1,\ldots,p \tag{6}$$

   using empirical search proposed by Pepe et al.: for each biomarker pair, for each value β of the 201 $\in [-1,1]$, the optimal cut-off point $(\hat{c}_\beta)$ that maximizes Youden index is selected. The final value chosen $(\hat{\beta})$ is the one with the highest Youden $(\hat{J}_\beta)$ obtained;
2. Once the pair of biomarkers and the parameter that maximizes the Youden index are chosen, this linear combination is considered as a single variable. For simplicity, suppose the linear combination $X_{ki1} + \beta_2 X_{ki2}$. Then, in the same way as point 1, the biomarker X_{kij} (of the remaining $p-2$s) and the β_3 parameter whose new linear combination maximize the Youden index are selected:

$$\hat{J}_{\beta_3} = \frac{\sum_{i=1}^{n_2} I\left((X_{2i1} + \beta_2 X_{2i2}) + \beta_3 X_{2ij} \leq \hat{c}_{\beta_3}\right)}{n_2} - \frac{\sum_{i=1}^{n_1} I\left((X_{1i1} + \beta_2 X_{1i2}) + \beta_3 X_{1ij} \leq \hat{c}_{\beta_3}\right)}{n_1} \qquad \beta_3 \in [-1,1], \ \forall j = 3,\ldots,p \tag{7}$$

$$\hat{J}_{\beta_3} = \frac{\sum_{i=1}^{n_2} I\left(\beta_3(X_{2i1} + \beta_2 X_{2i2}) + X_{2ij} \leq \hat{c}_{\beta_3}\right)}{n_2} - \frac{\sum_{i=1}^{n_1} I\left(\beta_3(X_{1i1} + \beta_2 X_{1i2}) + X_{1ij} \leq \hat{c}_{\beta_3}\right)}{n_1} \qquad \beta_3 \in [-1,1], \ \forall j = 3,\ldots,p \tag{8}$$

   Specifically, either the combination (7) or (8) that maximizes the Youden index $\hat{J}_{\beta_3}$ is selected. This new linear combination will be considered as a new variable in the next step;
3. The process (2) is repeated for the rest of biomarkers (i.e., $p-3$ times) until all of them are included in the model.

At each step, the maximum Youden index can be reached for more than one optimal linear combination. Our proposed algorithm considers each of these combinations and generates a branch to be explored by the algorithm. That is, it considers all ties at each stage and drags them forward until they are broken in the next steps (whenever possible) or until the end of the algorithm.

2.3. Yin and Tian's Stepwise Approach

The stepwise non-parametric approach with downward direction (SWD) introduced by Yin and Tian [27] is an adaptation of the step-down approach proposed by Kang et al. [10,18] for the Youden index maximization. As the stepwise approach previously described, the general idea is to introduce a new variable at each stage and find the combination of two variables that maximizes the Youden index using the empirical search for combination parameters proposed by Pepe et al. [13,16].

Unlike our proposed stepwise approach, where, in each step, a search is performed not only for the parameter β but also for the new biomarker that obtains the best linear combination, the approach proposed by Yin and Tian sets the biomarker that is entered in

each step, based on the values ordered from largest to smallest of the empirical Youden index, obtained for each biomarker as follows:

$$\hat{J}_j = \frac{\sum_{i=1}^{n_2} I\left(X_{2ij} \leq \hat{c}_j\right)}{n_2} - \frac{\sum_{i=1}^{n_1} I\left(X_{1ij} \leq \hat{c}_j\right)}{n_1} \qquad \forall j = 1, \ldots, p \tag{9}$$

Therefore, the approach is reduced to choosing, in each step, the parameter β whose linear combination achieves the highest Youden index. Another difference from our proposed stepwise algorithm is that the Yin and Tian approach does not handle ties but chooses only one combination from among the optimal ones at each step.

Therefore, the approach presented by Yin and Tian could be considered as a simpler particular case of our proposed stepwise approach, where the new biomarkers of each stage are fixed from the beginning and where the ties are not considered.

2.4. Min-Max Approach

The non-parametric min-max approach (MM) was proposed by Liu et al. [15]. The aim was to reduce the order of the linear combination by considering only two markers (maximum value and the minimum value of all the p biomarkers) and estimate only the parameter β of the linear combination that maximizes the AUC. Under this idea, the min-max approach can be adapted to maximize the Youden index with an expression as follows:

$$\hat{J}_\beta = \frac{\sum_{i=1}^{n_2} I\left(X_{2i,max} + \beta X_{2i,min} \leq \hat{c}_\beta\right)}{n_2} - \frac{\sum_{i=1}^{n_1} I\left(X_{1i,max} + \beta X_{1i,min} \leq \hat{c}_\beta\right)}{n_1} \tag{10}$$

where $X_{ki,max} = max_{1 \leq j \leq p}\left(X_{kij}\right)$ and $X_{ki,min} = min_{1 \leq j \leq p}\left(X_{kij}\right)$ for $k = 1, 2$ and each $i = 1, \ldots, n_k$, and $\beta \in [-1, 1]$, following Pepe et al's. [13,16] suggestion of the empirical search for the optimal value of β.

2.5. Logistic Regression

The logistic regression [31] (LR) (or logit regression) is a statistical model that models the probability of an event (disease or non-disease) given a set of independent variables through the logistics function. Assuming that the set of predictive independent variables for patient i is $\mathbf{X_i} = (X_{i1}, \ldots, X_{ip})^T$, the classification problem becomes the estimation of the parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$, such that:

$$\begin{aligned} P(\mathbf{Y_i} = 1 | \mathbf{X_i}) &= \frac{1}{1 + e^{-\boldsymbol{\beta}^T \mathbf{X_i}}} \\ &= \frac{e^{\boldsymbol{\beta}^T \mathbf{X_i}}}{1 + e^{\boldsymbol{\beta}^T \mathbf{X_i}}} \end{aligned} \tag{11}$$

and linear dependence:

$$log \frac{P(\mathbf{Y_i} = 1 | \mathbf{X_i})}{1 - P(\mathbf{Y_i} = 1 | \mathbf{X_i})} = \boldsymbol{\beta}^T \mathbf{X_i} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} \qquad \forall i = 1, \ldots, n_1 + n_2 \tag{12}$$

For the application of the logistic regression model, the R function glm() was used.

2.6. Parametric Approach under Multivariate Normality

The parametric approach to estimate the Youden index under multivariate normality (MVN) is based on the results presented by Schisterman and Perkins [32].

Suppose $\mathbf{X_k} \sim MVN(\mathbf{m_k}, \boldsymbol{\Sigma_k})$ and the single marker $\mathbf{Y_k} \sim N(\mu_k, \sigma_k^2)$, the result of linear combination is $\mathbf{Y_k} = \boldsymbol{\beta}^T \mathbf{X_k}$, where:

$$\mu_k = \boldsymbol{\beta}^T \mathbf{m_k}, \;\; \sigma_k = \sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma_k} \boldsymbol{\beta}} \tag{13}$$

for $k = 1, 2$ (disease and non-disease groups, respectively).

The formula for the Youden index and the optimal cut-off point differs depending on whether $\sigma_1^2 \neq \sigma_2^2$ (i.e., $\Sigma_1 \neq \Sigma_2$) or $\sigma_1^2 = \sigma_2^2$ (i.e., $\Sigma_1 = \Sigma_2$). Under the first scenario ($\Sigma_1 \neq \Sigma_2$), for $\mathbf{Y_k}$, the Youden index (J_β) and the optimal cut-off point (c_β) are expressed as follows:

$$J_\beta = \Phi\left(\frac{c_\beta - \mu_2}{\sigma_2}\right) - \Phi\left(\frac{c_\beta - \mu_1}{\sigma_1}\right), \quad c_\beta = \frac{\mu_1\sigma_2^2 - \mu_2\sigma_1^2 - \sigma_1\sigma_2\sqrt{(\mu_2 - \mu_1)^2 + (\sigma_2^2 - \sigma_1^2)\ln\frac{\sigma_2^2}{\sigma_1^2}}}{\sigma_2^2 - \sigma_1^2} \tag{14}$$

where Φ indicates the normal cumulative distribution function. Under the second one ($\Sigma_1 = \Sigma_2$), the expressions are the following:

$$J_\beta = 2\Phi\left(\frac{\mu_1 - \mu_2}{2\sqrt{\sigma_1^2}}\right) - 1, \quad c_\beta = \frac{\mu_1 + \mu_2}{2} \tag{15}$$

These formulations are also valid under Box–Cox-type transformations [33].

Note that the Youden index (J_β) is a continuous differentiable function with respect to the parameter vector $\boldsymbol{\beta}$ and its estimation ($\hat{J}_\beta$) can be numerically optimized from quasi-Newton algorithms. Specifically, the R package optimr() was used to estimate the parameter vector $\boldsymbol{\beta}$ from a initial parameter vector.

2.7. Non-Parametric Kernel Smoothing Approach

When no distributional hypothesis can be assumed, empirical distribution functions are often used, and their estimations can be performed using kernel-type approximations. In particular, a non-parametric Kernel Smoothing approach (*KS*) was applied in our study, whose estimation of the Youden index is as follows:

$$\begin{aligned}\hat{J}_\beta^{KS} &= \hat{F}_{\mathbf{Y_2}}^{KS}(\hat{c}_\beta^{KS}) - \hat{F}_{\mathbf{Y_1}}^{KS}(\hat{c}_\beta^{KS}) \\ &= \frac{1}{n_2}\sum_{i=1}^{n_2}\Phi\left(\frac{\hat{c}_\beta^{KS} - Y_{2i}}{h_{Y_2}}\right) - \frac{1}{n_1}\sum_{i=1}^{n_1}\Phi\left(\frac{\hat{c}_\beta^{KS} - Y_{1i}}{h_{Y_1}}\right) \\ &= \frac{1}{n_2}\sum_{i=1}^{n_2}\Phi\left(\frac{\hat{c}_\beta^{KS} - \boldsymbol{\beta}^T X_{2i}}{h_{Y_2}}\right) - \frac{1}{n_1}\sum_{i=1}^{n_1}\Phi\left(\frac{\hat{c}_\beta^{KS} - \boldsymbol{\beta}^T X_{1i}}{h_{Y_1}}\right)\end{aligned} \tag{16}$$

where the kernel function Φ is the normal cumulative distribution function and the general-purpose bandwidth h_{Y_k} [34–37] is:

$$h_{Y_k} = 0.9\min\left\{\frac{SD(Y_k), IQR(Y_k)}{1.34}\right\}n_k^{-0.2}, \text{ for } k = 1, 2,$$

where $SD(Y_k)$ and $IQR(Y_k)$ denote the standard deviation and the interquartile range of the combined marker Y_k, respectively.

Note that the Youden index ($\hat{J}_\beta$) is a continuous differentiable function with respect to the parameter vector $\boldsymbol{\beta}$ and c_β. As in the previous approach, the R package optimr() was used to numerically optimize the Youden index J_β^{KS} from an initial vector $\left(\hat{\boldsymbol{\beta}}^T, \hat{c}_\beta^{KS}\right)^T$. Thus, the estimated parameter vector $\boldsymbol{\beta}$ can be obtained.

2.8. Simulations

A wide range of simulated scenarios were explored in order to compare the performance of the algorithms. Four ($p = 4$) biomarkers were considered in each simulation scenario. Different joint and marginal distributions were considered, ranging from normal to non-normal distributions, with the aim of broadening the range for the evaluation and comparison of methods beyond normality.

A wide range of combinations were considered for the generation of simulated data following normal distributions: biomarkers with equal or different means (i.e., different capacity to discriminate between biomarkers) and independent or non-independent biomarkers, with negative or positive correlations with low, medium and high intensity, as well as the same and different covariances matrix for the group with disease and without disease. For each scenario, 1000 random samples from the underlying distribution were considered, with different sample sizes for the diseased and non-diseased population: $(n_1, n_2) = (10, 20), (30, 30), (50, 30), (50, 50), (100, 100), (500, 500)$. Each method was applied on each simulated dataset and the maximum Youden index for the optimal linear combination of biomarkers was obtained.

In terms of the scenarios of normal distributions, the null vector is considered in all scenarios as the mean vector of the non-diseased population ($m_2 = (0, 0, 0, 0)^T$). With respect to the diseased population (m_1), scenarios are explored with the same mean for each biomarker, as well as with mean vectors with different values. As for the covariance matrix, scenarios are analyzed with both the same covariance matrices for both populations and with different covariance matrices. For simplicity, the variance of each biomarker is set to be 1 in all cases and, therefore, the covariances equal to the correlations. The same correlation value for all pairs of biomarkers is assumed. Both positive and negative correlations are considered. Concerning negative correlations, the values $\rho = -0.3$ and $\rho = -0.1$ are considered. Regarding positive correlations, four types of correlations are assumed depending on the intensity: independence ($\rho = 0.0$), low ($\rho = 0.3$), medium ($\rho = 0.5$) and high ($\rho = 0.7$). Specifically, the following covariance matrices (Σ_1, Σ_2 for diseased and non-diseased population, respectively) are considered in the different scenarios: $\Sigma_1 = \Sigma_2 = I$ (independent biomarkers), $\Sigma_1 = \Sigma_2 = 0.7 \cdot I + 0.3 \cdot J$ (low correlation), $\Sigma_1 = \Sigma_2 = 0.5 \cdot I + 0.5 \cdot J$ (medium correlation), $\Sigma_1 = \Sigma_2 = 0.3 \cdot I + 0.7 \cdot J$ (high correlation) and $\Sigma_1 = 0.3 \cdot I + 0.7 \cdot J$, $\Sigma_2 = 0.7 \cdot I + 0.3 \cdot J$ (different correlations), where I is the identity matrix and J a matrix of all of them.

In terms of scenarios that do not follow a normal distribution, the following scenarios were considered: simulated data with different marginal distributions (multivariate chi-square/normal/gamma/exponential distributions via normal copula) and simulated data following the multivariate log-normal skewed distribution. The latter simulated data were generated from the normal scenario configurations and then exponentiated to obtain these multivariate log-normal observations.

2.9. Application in Clinical Diagnosis Cases

The analyzed methods were also applied to two real data examples related to clinical diagnosis cases. In particular, a Duchenne muscular dystrophy dataset and a prostate cancer dataset were analyzed through their respective biomarkers. Duchenne Muscular Dystrophy (DMD) is a progressive and recessive muscular disorder that is transmitted from a mother to her children. Percy et al. [38] analyzed the effectiveness in detecting the following four biomarkers of blood samples: serum creatine kinase (CK), haemopexin (H), pyruvate kinase (PK) and lactate dehydrogenase (LDH). The available data contain complete information on these four biomarkers of 67 women who are carriers of the progressive recessive disorder DMD and 127 women who are not carriers.

Prostate cancer is the second most common cancer in males worldwide after lung cancer [39] and it is therefore a matter of social and medical concern. The detection of clinically significant prostate cancer (Gleason score ≥ 7) through the combination of clinical characteristics and biomarkers has been an important line of study in recent years [40,41]. The data set used contains complete information on 71 people who were diagnosed with clinically significant prostate cancer and 398 with non-significant prostate cancer in 2016 at Miguel Servet University Hospital (Zaragoza, Spain) on the following four biomarkers: prostate-specific antigen (PSA), age, body mass index (BMI) and free PSA.

2.10. Validation

To analyze the performance of the compared algorithms in prediction scenarios, we validated built models for simulation and real data. For each scenario of simulated data, we built 100 models using small (50) and large (500) sample sizes, and then we validated these models by estimating the mean of the 100 Youden indexes calculated for new data simulated using the same setting of parameters and sample sizes. For real data, a 10-fold cross validation procedure was performed.

3. Results

This section first presents the results of the simulations for the training set. Then, the results of simulated scenarios for the validation data are presented. Finally, for a specific scenario, the time carried out in each of the methods is also presented in order to illustrate the computational cost of each one of them.

3.1. Simulations

Tables 1–10 show the results of the performance of each algorithm for each simulated data scenario. In particular, for each simulated dataset (1000 random samples), the mean and standard deviation (SD) of the empirical estimates of the Youden index of each biomarker are shown. In addition, for each method, the mean and standard deviation of the maximum Youden indexes obtained in each of the 1000 samples, as well as the probability of obtaining the highest Youden index, are presented. These results and conclusions drawn in terms of performance in the simulated scenarios are presented below.

3.1.1. Normal Distributions. Different Means and Equal Positive Correlations for Diseased and Non-Diseased Population

Tables 1–4 show the results obtained in the scenarios under multivariate normal distribution with mean vectors $m_1 = (0.2, 0.5, 1.0, 0.7)^T$ and $m_1 = (0.4, 1.0, 1.5, 1.2)^T$ and independent biomarkers, low correlations, medium correlations and high correlations, respectively.

The results in Table 1 show that our proposed stepwise method outperforms the rest of the methods in all scenarios and with a remarkable estimated probability of yielding the largest Youden index of 0.5 or more in most of them. It is followed by Yin and Tian's stepwise approach and the non-parametric kernel smoothing approach, which perform similarly in general. Logistic regression and the parametric approach under multivariate normality perform comparably in general. The min-max approach is the one with the worst results in such scenarios. The same conclusions are drawn from the results reported in Table 2.

Table 1. Normal distributions: Different means. Independence ($\Sigma_1 = \Sigma_2 = I$).

Size (n_1, n_2)	Mean (SD) Variables	Mean (SD)						Probability Greater than or Equal to Youden Index					
		SLM	SWD	MM	LR	MVN	KS	SLM	SWD	MM	LR	MVN	KS
					$m_1 = (0.2, 0.5, 1.0, 0.7)^T$								
(10, 20)	$\bar{x}$ = (0.2737, 0.3560, 0.5178, 0.4314)	**0.7782**	0.731	0.6352	0.6926	0.6937	0.7272	**0.5962**	0.1389	0.0588	0.0532	0.0355	0.1175
	σ = (0.1357, 0.1395, 0.1421, 0.1457)	(0.1022)	(0.1104)	(0.119)	(0.1346)	(0.1253)	(0.1179)						
(30, 30)	$\bar{x}$ = (0.2063, 0.3024, 0.4663, 0.3673)	**0.6737**	0.6402	0.5556	0.6057	0.6050	0.6395	**0.6480**	0.1350	0.0453	0.0221	0.0161	0.1335
	σ = (0.0943, 0.0991, 0.0990, 0.1053)	(0.0836)	(0.0876)	(0.0962)	(0.0957)	(0.0946)	(0.0891)						
(50, 30)	$\bar{x}$ = (0.1933, 0.2975, 0.4630, 0.3603)	**0.6434**	0.6278	0.5424	0.5895	0.5896	0.6203	**0.5806**	0.1756	0.0448	0.0179	0.0176	0.1636
	σ = (0.0846, 0.0937, 0.0894, 0.0898)	(0.0771)	(0.0778)	(0.0812)	(0.0839)	(0.0828)	(0.0806)						
(50, 50)	$\bar{x}$ = (0.1764, 0.2784, 0.4484, 0.3458)	**0.6219**	0.6005	0.5193	0.5693	0.5693	0.5998	**0.6586**	0.1428	0.0272	<0.01	0.0132	0.1495
	σ = (0.0736, 0.0789, 0.0774, 0.0796)	(0.0667)	(0.0702)	(0.0736)	(0.0732)	(0.0732)	(0.0701)						
(100, 100)	$\bar{x}$ = (0.1487, 0.2498, 0.4254, 0.3214)	**0.5734**	0.5623	0.4837	0.5412	0.5409	0.5620	**0.6174**	0.1543	0.0132	0.0171	0.0137	0.1844
	σ = (0.0533, 0.0590, 0.0560, 0.0590)	(0.0506)	(0.051)	(0.0526)	(0.0538)	(0.0537)	(0.0528)						
(500, 500)	$\bar{x}$ = (0.1062, 0.2185, 0.3991, 0.2925)	**0.5213**	0.5191	0.4447	0.5120	0.5119	0.5196	**0.4545**	0.1824	<0.01	0.0332	0.0317	0.2982
	σ = (0.0257, 0.0282, 0.0274, 0.0280)	(0.0257)	(0.0257)	(0.027)	(0.0262)	(0.0263)	(0.0258)						
					$m_1 = (0.4, 1.0, 1.5, 1.2)^T$								
(10, 20)	$\bar{x}$ = (0.3359, 0.5120, 0.6604, 0.5815)	**0.9134**	0.8783	0.8042	0.8771	0.8594	0.8786	**0.4822**	0.1128	0.0350	0.1913	0.0565	0.1221
	σ = (0.1413, 0.1386, 0.1275, 0.1402)	(0.074)	(0.0834)	(0.1013)	(0.0158)	(0.0958)	(0.0886)						
(30, 30)	$\bar{x}$ = (0.2699, 0.4695, 0.6172, 0.5299)	**0.8488**	0.8190	0.7463	0.8086	0.8044	0.8242	**0.5826**	0.1304	0.0319	0.0692	0.0420	0.1438
	σ = (0.0989, 0.0988, 0.0911, 0.1019)	(0.0645)	(0.0690)	(0.0787)	(0.0778)	(0.0750)	(0.0719)						
(50, 30)	$\bar{x}$ = (0.2586, 0.4636, 0.6133, 0.5218)	**0.8310**	0.8172	0.7400	0.8005	0.7960	0.8162	**0.5270**	0.1755	0.0372	0.0540	0.0362	0.1700
	σ = (0.0905, 0.0926, 0.0810, 0.0854)	(0.0598)	(0.0621)	(0.069)	(0.0676)	(0.0666)	(0.0633)						
(50, 50)	$\bar{x}$ = (0.2444, 0.4475, 0.6010, 0.5099)	**0.8144**	0.7972	0.7235	0.7840	0.7806	0.8007	**0.5841**	0.1403	0.0199	0.0436	0.0295	0.1826
	σ = (0.0791, 0.0796, 0.0695, 0.0772)	(0.0545)	(0.0569)	(0.0624)	(0.0598)	(0.0584)	(0.0573)						
(100, 100)	$\bar{x}$ = (0.2178, 0.4243, 0.5821, 0.4906)	**0.7827**	0.7728	0.6987	0.7620	0.7611	0.7749	**0.5701**	0.1690	<0.01	0.0262	0.0286	0.2036
	σ = (0.0570, 0.0579, 0.0526, 0.0562)	(0.04)	(0.0412)	(0.0456)	(0.0423)	(0.0423)	(0.0412)						
(500, 500)	$\bar{x}$ = (0.1809, 0.3997, 0.5604, 0.4673)	**0.7483**	0.7461	0.6692	0.7424	0.7422	0.7471	**0.4667**	0.1695	<0.01	0.0431	0.0359	0.2848
	σ = (0.0271, 0.0272, 0.0255, 0.0260)	(0.02)	(0.0199)	(0.0223)	(0.0201)	(0.0201)	(0.0199)						

Table 2. Normal distributions: Different means. Low correlation ($\Sigma_1 = \Sigma_2 = 0.7 \cdot I + 0.3 \cdot J$).

Size (n_1, n_2)	Mean (SD) Variables	Mean (SD)						Probability Greater than or Equal to Youden Index					
		SLM	SWD	MM	LR	MVN	KS	SLM	SWD	MM	LR	MVN	KS
					$m_1 = (0.2, 0.5, 1.0, 0.7)^T$								
(10, 20)	$\bar{x}$ = (0.2666, 0.3621, 0.5180, 0.4226)	**0.7348**	0.6811	0.5730	0.6380	0.6385	0.6755	**0.6163**	0.1309	0.0617	0.0404	0.0312	0.1195
	σ = (0.1378, 0.1439, 0.1377, 0.1427)	(0.1063)	(0.1158)	(0.1294)	(0.1389)	(0.1358)	(0.1266)						
(30, 30)	$\bar{x}$ = (0.2037, 0.3051, 0.4700, 0.3774)	**0.6228**	0.5873	0.4842	0.5472	0.5483	0.5844	**0.6869**	0.1276	0.0343	0.0103	0.0125	0.1284
	σ = (0.0951, 0.1029, 0.1002 ,0.1004)	(0.0848)	(0.0896)	(0.0948)	(0.0994)	(0.0982)	(0.0939)						
(50, 30)	$\bar{x}$ = (0.1963,0.2917,0.4606,0.3620)	**0.5905**	0.5701	0.4677	0.5278	0.5284	0.5628	**0.6378**	0.1491	0.0322	0.0151	0.0141	0.1517
	σ = (0.0841, 0.0888, 0.0899, 0.0903)	(0.0788)	(0.0787)	(0.0831)	(0.0868)	(0.0862)	(0.0852)						
(50, 50)	$\bar{x}$ = (0.1771, 0.2780, 0.4448, 0.3459)	**0.5641**	0.5380	0.4435	0.5030	0.5038	0.5354	**0.7324**	0.1176	0.0165	0.0123	0.0113	0.1098
	σ = (0.0751, 0.0798, 0.0778, 0.0805)	(0.0683)	(0.0707)	(0.0753)	(0.0752)	(0.0756)	(0.0717)						
(100, 100)	$\bar{x}$ = (0.1464, 0.2526, 0.4270, 0.3230)	**0.5148**	0.5012	0.4078	0.4770	0.4768	0.4984	**0.6986**	0.1268	<0.01	0.013	<0.01	0.1457
	σ = (0.0515, 0.0593, 0.0580, 0.0591)	(0.0546)	(0.0550)	(0.0542)	(0.0584)	(0.0585)	(0.0563)						
(500, 500)	$\bar{x}$ = (0.1063, 0.2178, 0.3980, 0.2923)	**0.4524**	0.449	0.3586	0.4404	0.4402	0.4488	**0.5629**	0.1693	<0.01	0.0234	0.0222	0.2221
	σ = (0.0262, 0.0282, 0.0278, 0.0278)	(0.0264)	(0.0265)	(0.0271)	(0.0269)	(0.0268)	(0.0265)						
					$m_1 = (0.4, 1.0, 1.5, 1.2)^T$								
(10, 20)	$\bar{x}$ = (0.3272, 0.5176, 0.6594, 0.5757)	**0.8558**	0.8128	0.7198	0.7941	0.7847	0.809	**0.5515**	0.1334	0.0565	0.1059	0.0445	0.1081
	σ = (0.1436, 0.1440, 0.1291, 0.1388)	(0.0907)	(0.1001)	(0.1174)	(0.1256)	(0.1137)	(0.1056)						
(30, 30)	$\bar{x}$ = (0.2685, 0.4690, 0.6182, 0.5362)	**0.7751**	0.7433	0.6514	0.7196	0.7175	0.7433	**0.6647**	0.1203	0.0238	0.0329	0.0271	0.1313
	σ = (0.1006, 0.1013, 0.0920, 0.0949)	(0.0742)	(0.0772)	(0.0892)	(0.0852)	(0.0841)	(0.082)						
(50,30)	$\bar{x}$ = (0.2629, 0.4602, 0.6113, 0.5246)	**0.7515**	0.7343	0.6393	0.7057	0.7047	0.7291	**0.643**	0.1586	0.0228	0.0243	0.0204	0.1308
	σ = (0.0891, 0.0909, 0.0813, 0.0870)	(0.0681)	(0.0705)	(0.078)	(0.0779)	(0.0762)	(0.0733)						
(50, 50)	$\bar{x}$ = (0.2441, 0.4483, 0.5975, 0.5108)	**0.7307**	0.7107	0.6204	0.6883	0.6870	0.7109	**0.6652**	0.1333	0.0168	0.0217	0.021	0.142
	σ = (0.0793, 0.0795, 0.0729, 0.0761)	(0.0609)	(0.0646)	(0.0702)	(0.0675)	(0.0673)	(0.0648)						
(100, 100)	$\bar{x}$ = (0.2160, 0.4258, 0.5829, 0.4909)	**0.6934**	0.6818	0.5922	0.6642	0.6641	0.6814	**0.655**	0.1473	<0.01	0.0229	0.0229	0.1485
	σ = (0.0537, 0.0578, 0.0529, 0.0567)	(0.0488)	(0.0488)	(0.0496)	(0.0525)	(0.0516)	(0.0492)						
(500, 500)	$\bar{x}$ = (0.1803, 0.3987, 0.5594, 0.4663)	**0.6453**	0.643	0.5543	0.6379	0.6378	0.6436	**0.4826**	0.1741	<0.01	0.0363	0.0354	0.2717
	σ = (0.0277, 0.0274, 0.0252, 0.0266)	(0.023)	(0.023)	(0.0244)	(0.0237)	(0.0238)	(0.0229)						

The results reported in Table 3 show that, in general, our proposed stepwise method dominates over the rest of the algorithms. The non-parametric kernel smoothing approach slightly outperforms Yin and Tian's approach in terms of the average Youden index, and even for large sample sizes ($n_1 = n_2 = 500$), its mean Youden index is even slightly higher than that of our proposed stepwise method. This behaviour is accentuated in the scenarios of Table 4, where Yian and Tian's stepwise approach performs significantly worse than the non-parametric kernel smoothing approach, and even their average values are lower than those achieved by logistic regression or the parametric approach under multivariate normality.

Table 3. Normal distributions: Different means. Medium correlation ($\Sigma_1 = \Sigma_2 = 0.5 \cdot I + 0.5 \cdot J$).

Size (n_1, n_2)	Mean (SD) Variables	Mean (SD) SLM	SWD	MM	LR	MVN	KS	Probability Greater than or Equal to Youden Index SLM	SWD	MM	LR	MVN	KS
					$m_1 = (0.2, 0.5, 1.0, 0.7)^T$								
(10, 20)	$\bar{x}$ = (0.2713, 0.3657, 0.5268, 0.428)	**0.7314**	0.667	0.5534	0.6404	0.6432	0.674	**0.4883**	0.1566	0.0399	0.0496	0.0804	0.1851
	σ = (0.1353, 0.1477, 0.1423, 0.1427)	(0.1093)	(0.1181)	(0.1271)	(0.1387)	(0.1356)	(0.1306)						
(30, 30)	$\bar{x}$ = (0.2030, 0.3020, 0.4665, 0.3668)	**0.62**	0.5647	0.4598	0.548	0.5433	0.5777	**0.5478**	0.1417	0.0118	0.0244	0.0572	0.2171
	σ = (0.0933, 0.1014, 0.0991, 0.1008)	(0.0847)	(0.0882)	(0.0945)	(0.0972)	(0.0952)	(0.0921)						
(50, 30)	$\bar{x}$ = (0.1931, 0.2928, 0.4581, 0.3632)	**0.5874**	0.5568	0.4415	0.5288	0.5297	0.5631	**0.4942**	0.1885	<0.01	0.0295	0.0282	0.2543
	σ = (0.0847, 0.0908, 0.0883 ,0.0905)	(0.0769)	(0.0764)	(0.0811)	(0.0856)	(0.0842)	(0.0831)						
(50, 50)	$\bar{x}$ = (0.1753, 0.2761, 0.4472, 0.3495)	**0.5637**	0.5274	0.4187	0.5096	0.5094	0.5401	**0.5076**	0.1588	<0.01	0.0247	0.0502	0.2528
	σ = (0.0742, 0.0798, 0.0787, 0.0791)	(0.0677)	(0.0717)	(0.0738)	(0.0755)	(0.0746)	(0.0738)						
(100, 100)	$\bar{x}$ = (0.1449, 0.2487, 0.4250, 0.3236)	**0.5114**	0.4907	0.3806	0.4766	0.4766	0.5005	**0.4238**	0.3977	<0.01	0.0152	<0.01	0.1548
	σ = (0.0520, 0.0578, 0.0591, 0.0595)	(0.0539)	(0.0551)	(0.056)	(0.0583)	(0.0584)	(0.0566)						
(500, 500)	$\bar{x}$ = (0.1063, 0.2187, 0.3994, 0.2921)	0.4511	0.443	0.3325	0.4429	0.4429	**0.4516**	0.3103	**0.3948**	<0.01	0.0237	0.0204	0.2508
	σ = (0.0261, 0.0280, 0.0264, 0.0279)	(0.0256)	(0.0276)	(0.0272)	(0.0265)	(0.0265)	(0.0256)						
					$m_1 = (0.4, 1.0, 1.5, 1.2)^T$								
(10, 20)	$\bar{x}$ = (0.3339, 0.5172, 0.6658, 0.5790)	**0.844**	0.7906	0.6879	0.7838	0.7736	0.7996	**0.4078**	0.1614	0.0258	0.1348	0.0884	0.1817
	σ = (0.1406, 0.1442, 0.1287, 0.1370)	(0.0934)	(0.1050)	(0.1184)	(0.1288)	(0.1167)	(0.1107)						
(30, 30)	$\bar{x}$ = (0.2685, 0.4669, 0.6160, 0.5254)	**0.7609**	0.7139	0.6157	0.7084	0.7007	0.7294	**0.5169**	0.1461	0.0115	0.0512	0.0628	0.2116
	σ = (0.0998, 0.1025, 0.0945, 0.0954)	(0.0766)	(0.0801)	(0.0906)	(0.0876)	(0.087)	(0.0804)						
(50, 30)	$\bar{x}$ = (0.2580, 0.4621, 0.6105, 0.5264)	**0.7348**	0.7101	0.6034	0.6945	0.6929	0.7177	**0.4785**	0.1823	<0.01	0.0643	0.0438	0.2257
	σ = (0.0887, 0.0908, 0.0807, 0.0862)	(0.0686)	(0.0707)	(0.0782)	(0.0773)	(0.0743)	(0.0721)						
(50, 50)	$\bar{x}$ = (0.2428, 0.4477, 0.5994, 0.5124)	**0.7162**	0.6874	0.5831	0.6778	0.6762	0.7013	**0.4817**	0.1485	<0.01	0.0496	0.0575	0.2574
	σ = (0.0782, 0.0810, 0.0723, 0.0769)	(0.0620)	(0.0649)	(0.0715)	(0.0697)	(0.0665)	(0.0639)						
(100, 100)	$\bar{x}$ = (0.2146, 0.4235, 0.5816, 0.4916)	**0.6755**	0.6572	0.5535	0.6519	0.6513	0.6684	**0.5904**	0.083	<0.01	0.04	0.0382	0.2475
	σ = (0.0554, 0.0572, 0.0550, 0.0580)	(0.0480)	(0.0493)	(0.0524)	(0.0510)	(0.0509)	(0.0494)						
(500, 500)	$\bar{x}$ = (0.1808, 0.3990, 0.5606, 0.4663)	0.6286	0.6239	0.5165	0.6258	0.6255	**0.6317**	0.379	0.1055	<0.01	0.0561	0.0563	**0.403**
	σ = (0.0270, 0.0277, 0.0244, 0.0260)	(0.0236)	(0.0242)	(0.0251)	(0.0232)	(0.0232)	(0.023)						

Table 4. Normal distributions: Different means. High correlation ($\Sigma_1 = \Sigma_2 = 0.3 \cdot I + 0.7 \cdot J$).

Size (n_1, n_2)	Mean (SD) Variables	Mean (SD) SLM	SWD	MM	LR	MVN	KS	Probability Greater than or Equal to Youden Index SLM	SWD	MM	LR	MVN	KS
					$m_1 = (0.2, 0.5, 1.0, 0.7)^T$								
(10, 20)	$\bar{x}$ = (0.2708, 0.3672, 0.5202, 0.4299)	**0.754**	0.6634	0.5546	0.6702	0.6704	0.6962	**0.5549**	0.1035	0.0211	0.0879	0.0684	0.1642
	σ = (0.1361, 0.1398, 0.1435, 0.1447)	(0.1036)	(0.119)	(0.1268)	(0.1418)	(0.1383)	(0.1317)						
(30, 30)	$\bar{x}$ = (0.2005, 0.3040, 0.4663, 0.3703)	**0.6514**	0.5668	0.4559	0.5844	0.5834	0.6166	**0.5812**	0.0826	0.0116	0.0521	0.0397	0.2329
	σ = (0.0936, 0.1025, 0.1045, 0.1038)	(0.0817)	(0.0937)	(0.0946)	(0.0981)	(0.0972)	(0.0940)						
(50, 30)	$\bar{x}$ = (0.1951, 0.2932, 0.4625, 0.3591)	**0.6175**	0.5628	0.4401	0.5689	0.5690	0.6000	**0.5288**	0.1146	<0.01	0.0414	0.0324	0.2768
	σ = (0.0815, 0.0877, 0.0893, 0.0914)	(0.0779)	(0.0834)	(0.0817)	(0.0882)	(0.0874)	(0.0859)						
(50, 50)	$\bar{x}$ = (0.1781, 0.2779, 0.4479, 0.3450)	**0.5983**	0.5325	0.4166	0.5527	0.5521	0.5805	**0.5430**	0.0799	<0.01	0.0454	0.0431	0.2858
	σ = (0.0735, 0.0787, 0.0800, 0.0806)	(0.0671)	(0.0754)	(0.0747)	(0.0782)	(0.0774)	(0.0741)						
(100, 100)	$\bar{x}$ = (0.1461, 0.2526, 0.4243, 0.3207)	**0.5472**	0.4969	0.3773	0.5203	0.5202	**0.5413**	0.4680	0.2348	<0.01	0.0173	0.0222	0.2577
	σ = (0.0534, 0.0597, 0.0566, 0.0606)	(0.0515)	(0.0567)	(0.0548)	(0.0547)	(0.0547)	(0.0530)						
(500, 500)	$\bar{x}$ = (0.1057, 0.2177, 0.3992, 0.2928)	0.4949	0.4587	0.3313	0.4893	0.4892	**0.4972**	0.3588	0.1517	<0.01	0.0418	0.0358	**0.4118**
	σ = (0.0267, 0.0281, 0.0276, 0.0278)	(0.0257)	(0.0308)	(0.0274)	(0.0260)	(0.0259)	(0.0255)						
					$m_1 = (0.4, 1.0, 1.5, 1.2)^T$								
(10, 20)	$\bar{x}$ = (0.3322, 0.5204, 0.6594, 0.5826)	**0.8612**	0.7759	0.6874	0.8091	0.7979	0.8204	**0.5229**	0.0699	0.0183	0.1656	0.0730	0.1503
	σ = (0.1426, 0.1414, 0.1301, 0.1402)	(0.0882)	(0.1107)	(0.1232)	(0.1266)	(0.1156)	(0.1099)						
(30, 30)	$\bar{x}$ = (0.2669, 0.4678, 0.6169, 0.5312)	**0.7842**	0.7095	0.6146	0.7393	0.7369	0.7607	**0.5373**	0.0759	<0.01	0.0856	0.0629	0.2325
	σ = (0.0982, 0.0999, 0.0961, 0.1011)	(0.0702)	(0.0854)	(0.0884)	(0.0835)	(0.0821)	(0.0782)						
(50,30)	$\bar{x}$ = (0.2620, 0.4584, 0.6136, 0.5234)	**0.7596**	0.7128	0.6007	0.7304	0.7283	0.7514	**0.4932**	0.0898	<0.01	0.0842	0.0564	0.2714
	σ = (0.0865, 0.0865, 0.0817, 0.0866)	(0.0682)	(0.0722)	(0.078)	(0.0760)	(0.0735)	(0.0696)						
(50, 50)	$\bar{x}$ = (0.2458, 0.4465, 0.6006, 0.5098)	**0.7394**	0.6899	0.5806	0.7165	0.7149	0.7361	**0.4764**	0.0643	<0.01	0.0650	0.0637	0.3282
	σ = (0.0776, 0.0776, 0.0756, 0.0769)	(0.0612)	(0.0717)	(0.0729)	(0.0676)	(0.0663)	(0.0638)						
(100, 100)	$\bar{x}$ = (0.2155, 0.4256, 0.5809, 0.4890)	0.7018	0.6658	0.5523	0.6898	0.6886	**0.7036**	**0.4592**	0.0528	<0.01	0.0712	0.0544	0.3625
	σ = (0.0562, 0.0606, 0.0520, 0.0570)	(0.042)	(0.0525)	(0.0502)	(0.0469)	(0.047)	(0.0458)						
(500, 500)	$\bar{x}$ = (0.1802, 0.3992, 0.5603, 0.4668)	0.6588	0.6451	0.5148	0.6643	0.6643	**0.6701**	0.1175	0.2240	<0.01	0.0793	0.0653	**0.5138**
	σ = (0.0280, 0.0267, 0.0250, 0.0256)	(0.0226)	(0.0270)	(0.0253)	(0.0222)	(0.0222)	(0.0220)						

Given the reported results of these simulations, it could be concluded that, in scenarios of multivariate normal distributions with different means and equal positive correlations, our proposed stepwise method dominates generally over the rest of the algorithms, followed by the non-parametric kernel smoothing approach and Yin and Tian's stepwise approach, with the former being better in scenarios of higher correlations.

3.1.2. Normal Distributions. Different Means and Unequal Positive Correlations for Diseased and Non-Diseased Population

Table 5 shows the results obtained in the scenarios under multivariate normal distribution with mean vectors $m_1 = (0.2, 0.5, 1.0, 0.7)^T$ and $m_1 = (0.4, 1.0, 1.5, 1.2)^T$ and different correlations for the diseased and non-diseased populations ($\Sigma_1 = 0.3 \cdot I + 0.7 \cdot J$, $\Sigma_2 = 0.7 \cdot I + 0.3 \cdot J$). The results indicate that our proposed stepwise approach outperforms the other algorithms in most scenarios. It is followed by the non-parametric kernel smoothing approach and Yin and Tian's stepwise approach. The min-max approach is the worst

performer. Logistic regression and the parametric approach under multivariate normality performed comparably in general.

Table 5. Normal distributions: Different means. Different correlation ($\Sigma_1 = 0.3 \cdot I + 0.7 \cdot J$, $\Sigma_2 = 0.7 \cdot I + 0.3 \cdot J$).

Size (n_1, n_2)	Mean (SD) Variables	Mean (SD)						Probability Greater than or Equal to Youden Index					
		SLM	SWD	MM	LR	MVN	KS	SLM	SWD	MM	LR	MVN	KS
					$m_1 = (0.2, 0.5, 1.0, 0.7)^T$								
(10, 20)	$\bar{x}$ = (0.2706, 0.3608, 0.5172, 0.4272)	**0.736**	0.6644	0.5952	0.637	0.6408	0.669	**0.4899**	0.1413	0.0987	0.0536	0.0679	0.1486
	σ = (0.1367, 0.1405, 0.1453, 0.1456)	(0.0998)	(0.116)	(0.1258)	(0.1328)	(0.1338)	(0.1292)						
(30, 30)	$\bar{x}$ = (0.2023, 0.3027, 0.4703, 0.3748)	**0.6268**	0.5745	0.5086	0.5512	0.5527	0.5871	**0.5236**	0.1399	0.0877	0.0244	0.038	0.1864
	σ = (0.0948, 0.1002, 0.0979, 0.1003)	(0.0825)	(0.0869)	(0.0905)	(0.0973)	(0.0928)	(0.0896)						
(50, 30)	$\bar{x}$ = (0.1957, 0.2895, 0.4586, 0.3606)	**0.5986**	0.567	0.4958	0.5373	0.5383	0.5717	**0.4958**	0.1717	0.0993	0.0327	0.0282	0.1723
	σ = (0.0832, 0.0925, 0.0862, 0.0918)	(0.0794)	(0.0775)	(0.0827)	(0.0907)	(0.0868)	(0.0856)						
(50, 50)	$\bar{x}$ = (0.1785, 0.2736, 0.4439, 0.3447)	**0.5727**	0.5297	0.4708	0.5096	0.5129	0.5426	**0.5447**	0.1275	0.0869	0.0149	0.0376	0.1883
	σ = (0.0735, 0.0793, 0.077, 0.0831)	(0.0669)	(0.0719)	(0.0745)	(0.0772)	(0.0763)	(0.0741)						
(100, 100)	$\bar{x}$ = (0.1465, 0.2526, 0.4273, 0.3225)	**0.5198**	0.4946	0.4378	0.482	0.4835	0.508	**0.4783**	0.3068	0.0648	0.0128	0.0113	0.1258
	σ = (0.052, 0.0572, 0.0576, 0.0598)	(0.052)	(0.0539)	(0.0557)	(0.0557)	(0.0556)	(0.0529)						
(500, 500)	$\bar{x}$ = (0.1063, 0.2181, 0.3994, 0.2929)	**0.4576**	0.4482	0.3916	0.4446	0.4464	0.4565	**0.3803**	0.3227	<0.01	0.0122	0.0187	0.2593
	σ = (0.0264, 0.0271, 0.0272, 0.0276)	(0.0256)	(0.0264)	(0.0271)	(0.0271)	(0.0268)	(0.026)						
					$m_1 = (0.4, 1.0, 1.5, 1.2)^T$								
(10, 20)	$\bar{x}$ = (0.3328, 0.5141, 0.6583, 0.5798)	**0.8422**	0.7851	0.6824	0.7763	0.7697	0.7934	**0.4419**	0.1638	0.0412	0.1108	0.0785	0.1637
	σ = (0.1429, 0.1394, 0.1338, 0.1404)	(0.091)	(0.1103)	(0.1244)	(0.1259)	(0.1146)	(0.1113)						
(30, 30)	$\bar{x}$ = (0.2664, 0.4658, 0.6184, 0.5334)	**0.7628**	0.7199	0.6068	0.7102	0.7055	0.7333	**0.5186**	0.1604	0.013	0.0512	0.056	0.2007
	σ = (0.1005, 0.0991, 0.0921, 0.0971)	(0.0752)	(0.0803)	(0.089)	(0.0869)	(0.0846)	(0.0802)						
(50, 30)	$\bar{x}$ = (0.2627, 0.4576, 0.6125, 0.5232)	**0.7403**	0.7154	0.5945	0.6961	0.6947	0.7219	**0.5055**	0.2028	0.0115	0.0428	0.0379	0.1996
	σ = (0.0894, 0.0924, 0.0783, 0.088)	(0.0714)	(0.0713)	(0.0773)	(0.0827)	(0.0767)	(0.0722)						
(50, 50)	$\bar{x}$ = (0.2455, 0.4442, 0.5970, 0.5088)	**0.7176**	0.687	0.5735	0.6765	0.6759	0.6998	**0.5134**	0.1669	<0.01	0.048	0.045	0.2182
	σ = (0.0782, 0.0794, 0.0700, 0.0784)	(0.0617)	(0.0668)	(0.0713)	(0.0706)	(0.0676)	(0.0649)						
(100, 100)	$\bar{x}$ = (0.2150, 0.4275, 0.5840, 0.4909)	**0.6804**	0.661	0.5473	0.6543	0.6545	0.6724	**0.4505**	0.3459	<0.01	0.0262	0.021	0.1555
	σ = (0.0546, 0.0565, 0.0523, 0.0558)	(0.0477)	(0.0486)	(0.0537)	(0.0512)	(0.0509)	(0.0488)						
(500, 500)	$\bar{x}$ = (0.1807, 0.3992, 0.5600, 0.4667)	0.6309	0.6258	0.5091	0.6255	0.6258	**0.6326**	0.2942	**0.3528**	<0.01	0.0317	0.0346	0.2868
	σ = (0.0278, 0.0270, 0.0251, 0.0258)	(0.0233)	(0.0237)	(0.0256)	(0.024)	(0.0239)	(0.0235)						

3.1.3. Normal Distributions. Different Means and Equal Negative Correlations for Diseased and Non-Diseased Population

Tables 6 and 7 show the results obtained in the scenarios under multivariate normal distribution with mean vectors $m_1 = (0.2, 0.5, 1.0, 0.7)^T$ and $m_1 = (0.4, 1.0, 1.5, 1.2)^T$ and equal negative correlations ($\rho = -0.1, -0.3$, respectively).

Table 6. Normal distributions: Different means. Negative correlation (−0.1).

Size (n_1, n_2)	Mean (SD) Variables	Mean (SD)						Probability Greater than or Equal to Youden Index					
		SLM	SWD	MM	LR	MVN	KS	SLM	SWD	MM	LR	MVN	KS
					$m_1 = (0.2, 0.5, 1.0, 0.7)^T$								
(10, 20)	$\bar{x}$ = (0.2651, 0.3652, 0.5164, 0.4329)	**0.8127**	0.7625	0.678	0.7395	0.7356	0.7664	**0.4308**	0.1033	0.0474	0.1301	0.08	0.2084
	σ = (0.1298, 0.1462, 0.1450,0.1438)	(0.0967)	(0.105)	(0.1216)	(0.1334)	(0.1259)	(0.1163)						
(30, 30)	$\bar{x}$ = (0.2064, 0.3013, 0.4704, 0.3718)	**0.7108**	0.6747	0.5997	0.6603	0.6599	0.6905	**0.4499**	0.0831	0.0309	0.0744	0.0694	0.2924
	σ = (0.0949, 0.0980, 0.1005, 0.1040)	(0.0799)	(0.0823)	(0.0893)	(0.092)	(0.0904)	(0.0847)						
(50, 30)	$\bar{x}$ = (0.1947, 0.2934, 0.4588, 0.3593)	**0.6862**	0.6681	0.5863	0.6422	0.6413	0.6692	**0.4088**	0.132	0.0245	0.058	0.0495	0.3272
	σ = (0.0836, 0.0880, 0.0879, 0.0926)	(0.072)	(0.0742)	(0.0799)	(0.0829)	(0.0801)	(0.0788)						
(50, 50)	$\bar{x}$ = (0.1755, 0.2784, 0.4466, 0.3465)	**0.6667**	0.6419	0.568	0.6297	0.6288	0.654	**0.4187**	0.1011	0.0219	0.065	0.0756	0.3177
	σ = (0.0717, 0.0792, 0.0787, 0.0800)	(0.0622)	(0.0672)	(0.0721)	(0.0716)	(0.0704)	(0.0676)						
(100, 100)	$\bar{x}$ = (0.1445, 0.2522, 0.4264, 0.3212)	**0.6252**	0.6135	0.5321	0.6054	0.6053	0.6229	**0.3913**	0.1198	<0.01	0.05	0.0572	0.3759
	σ = (0.0531, 0.0578, 0.0579, 0.0585)	(0.0502)	(0.0507)	(0.0534)	(0.0515)	(0.0516)	(0.05)						
(500, 500)	$\bar{x}$ = (0.1068, 0.2180, 0.3989, 0.2923)	0.5784	0.5761	0.4947	0.5734	0.5735	**0.5804**	0.3137	0.1578	<0.01	0.0484	0.0567	**0.4234**
	σ = (0.0263, 0.0277, 0.0275, 0.0273)	(0.0237)	(0.0239)	(0.0251)	(0.0243)	(0.0244)	(0.0238)						
					$m_1 = (0.4, 1.0, 1.5, 1.2)^T$								
(10, 20)	$\bar{x}$ = (0.3266, 0.5185, 0.6594, 0.5868)	**0.9466**	0.9098	0.8538	0.9296	0.9078	0.922	**0.3156**	0.1426	0.056	0.267	0.0912	0.1276
	σ = (0.1344, 0.1434, 0.1322, 0.1347)	(0.0619)	(0.0752)	(0.0916)	(0.0898)	(0.0825)	(0.0757)						
(30, 30)	$\bar{x}$ = (0.2701, 0.4666, 0.6203, 0.5327)	**0.8952**	0.8625	0.8046	0.8737	0.8652	0.8835	**0.3993**	0.1047	0.026	0.168	0.0824	0.2195
	σ = (0.0986, 0.0999, 0.0926, 0.1011)	(0.056)	(0.0622)	(0.0697)	(0.0655)	(0.0625)	(0.0581)						
(50, 30)	$\bar{x}$ = (0.2602, 0.4611, 0.6101, 0.5226)	**0.8806**	0.8672	0.7964	0.8604	0.8545	0.8714	**0.3853**	0.1296	0.0278	0.1445	0.0751	0.2376
	σ = (0.0888, 0.0855, 0.0793, 0.0905)	(0.0516)	(0.0533)	(0.0644)	(0.0609)	(0.0578)	(0.055						
(50, 50)	$\bar{x}$ = (0.2427, 0.4476, 0.6006, 0.5110)	**0.8677**	0.8508	0.7835	0.8503	0.8454	0.8616	**0.381**	0.1215	0.0217	0.1144	0.0651	0.2963
	σ = (0.0757, 0.0791, 0.0731, 0.0765)	(0.0448)	(0.0477)	(0.0574)	(0.0506)	(0.0485)	(0.0476)						
(100, 100)	$\bar{x}$ = (0.2138, 0.4268, 0.5832, 0.4905)	**0.8442**	0.8342	0.757	0.8329	0.831	0.8427	**0.3815**	0.1121	<0.01	0.0856	0.0672	0.351
	σ = (0.0573, 0.0578, 0.0533, 0.0558)	(0.0367)	(0.0379)	(0.0422)	(0.037)	(0.0367)	(0.0356)						
(500, 500)	$\bar{x}$ = (0.1813, 0.3989, 0.5598, 0.4665)	**0.8152**	0.8127	0.7294	0.8108	0.8105	0.8145	**0.3765**	0.1367	<0.01	0.0724	0.0619	0.3525
	σ = (0.0273, 0.0270, 0.0255, 0.0255)	(0.0180)	(0.0182)	(0.0208)	(0.0176)	(0.0175)	(0.0173)						

Table 7. Normal distributions: Different means. Negative correlation (−0.3).

Size (n_1, n_2)	Mean (SD) Variables	Mean (SD)						Probability Greater than or Equal to Youden Index					
		SLM	SWD	MM	LR	MVN	KS	SLM	SWD	MM	LR	MVN	KS
					$m_1 = (0.2, 0.5, 1.0, 0.7)^T$								
(10, 20)	$\bar{x}$ = (0.2624, 0.3591, 0.5238, 0.4281) σ = (0.1354, 0.1447, 0.1413, 0.1425)	0.976 (0.0468)	0.8938 (0.0964)	0.845 (0.0935)	**0.9963** (0.0189)	0.9874 (0.0289)	0.9902 (0.0258)	0.2058	0.0617	0.0166	**0.2782**	0.2115	0.2261
(30, 30)	$\bar{x}$ = (0.2057, 0.3046, 0.4707, 0.3672) σ = (0.0942, 0.1007, 0.1004, 0.1008)	0.9589 (0.0472)	0.8946 (0.0852)	0.7906 (0.0717)	**0.9875** (0.0254)	0.9746 (0.027)	0.9821 (0.0243)	0.1978	0.0499	<0.01	**0.3625**	0.15661	0.2331
(50, 30)	$\bar{x}$ = (0.1967, 0.2948, 0.4565, 0.3621) σ = (0.0868, 0.0899, 0.0933, 0.0903)	0.9405 (0.0577)	0.9125 (0.0726)	0.7819 (0.0674)	**0.9835** (0.0267)	0.9725 (0.0269)	0.9802 (0.0243)	0.1498	0.0712	<0.01	**0.3874**	0.1407	0.2504
(50, 50)	$\bar{x}$ = (0.1769, 0.2795, 0.4447, 0.3482) σ = (0.0754, 0.0805, 0.0780, 0.0789)	0.9471 (0.0447)	0.9052 (0.0718)	0.7678 (0.0583)	**0.9775** (0.0255)	0.9668 (0.0243)	0.975 (0.0218)	0.192	0.0609	<0.01	**0.3861**	0.1163	0.2447
(100, 100)	$\bar{x}$ = (0.1435, 0.2540, 0.4258, 0.3240) σ = (0.0543, 0.0544, 0.0591, 0.0595)	0.9421 (0.0345)	0.9178 (0.0503)	0.7469 (0.0431)	0.9652 (0.0202)	0.9606 (0.019)	**0.9674** (0.0177)	0.1053	0.1346	<0.01	0.2778	0.1168	**0.3655**
(500, 500)	$\bar{x}$ = (0.1067, 0.2173, 0.3981, 0.2919) σ = (0.0270, 0.0284, 0.0280, 0.0276)	0.9365 (0.0195)	0.9309 (0.0217)	0.7161 (0.0215)	0.9509 (0.0096)	0.9502 (0.0097)	**0.9531** (0.0093)	0.1671	0.0586	<0.01	0.1500	0.1057	**0.5185**
					$m_1 = (0.4, 1.0, 1.5, 1.2)^T$								
(10, 20)	$\bar{x}$ = (0.3223, 0.5169, 0.6668, 0.5786) σ = (0.1401, 0.1425, 0.1279, 0.1364)	0.9998 (0.0034)	0.9876 (0.0311)	0.9734 (0.0413)	**1.0000** (0.00000)	**1.0000** (0.0000)	**1.0000** (0.0000)	0.1838	0.1479	0.1131	**0.1851**	**0.1851**	**0.1851**
(30, 30)	$\bar{x}$ = (0.2701, 0.4693, 0.6214, 0.5251) σ = (0.0990, 0.1018, 0.0922, 0.0968)	0.9997 (0.0031)	0.9891 (0.0249)	0.9519 (0.0384)	**1.0000** (0.0000)	**1.0000** (0.0000)	0.9999 (0.0015)	0.201	0.1504	0.0378	**0.2037**	**0.2037**	0.2032
(50, 30)	$\bar{x}$ = (0.2631, 0.4635, 0.6084, 0.5255 σ = (0.0933, 0.0889, 0.0858, 0.0867)	0.999 (0.0064)	0.9958 (0.0132)	0.9503 (0.0355)	1.0000 (0.0000)	**1.0000** (0.0006)	**1.0000** (0.0006)	0.1952	0.1699	0.0217	**0.2046**	0.2043	0.2043
(50, 50)	$\bar{x}$ = (0.2436, 0.4498, 0.5975, 0.5118) σ = (0.0795, 0.0802, 0.0713, 0.0768)	0.9995 (0.0033)	0.9953 (0.013)	0.9431 (0.0322)	**1.0000** (0.0000)	0.9999 (0.0015)	**1.0000** (0.0000)	0.2018	0.1672	<0.01	**0.2082**	0.2066	**0.2082**
(100, 100)	$\bar{x}$ = (0.2131, 0.4277, 0.5819, 0.4917) σ = (0.0578, 0.0541, 0.0543, 0.0577)	0.9996 (0.0022)	0.9981 (0.0057)	0.933 (0.0243)	**1.0000** (0.0000)	0.9999 (0.001)	**1.0000** (0.0000)	0.1985	0.1723	<0.01	**0.2107**	0.2077	**0.2107**
(500, 500)	$\bar{x}$ = (0.1812, 0.3992, 0.5598, 0.4662) σ = (0.0278, 0.0277, 0.0255, 0.0262)	0.9995 (0.0011)	0.9992 (0.0016)	0.916 (0.0121)	**0.9999** (0.0005)	0.9996 (0.0009)	0.9998 (0.0006)	0.1821	0.1581	<0.01	**0.247**	0.1911	0.2217

The same conclusions as in the previous tables can be deduced from Table 6 for the mean vector scenario $m_1 = (0.2, 0.5, 1.0, 0.7)^T$, globally. The results show that our proposed stepwise approach, in general, outperforms over the other algorithms. After it, the non-parametric kernel smoothing approach and Yin and Tian's stepwise approach are the best performers, the results of the former being slightly better than those of the latter. Logistic regression and the parametric approach under multivariate normality conditions obtain similar results. The min-max approach is the worst performer.

However, the results provided by the simulated data with mean vector $m_1 = (0.4, 1.0, 1.5, 1.2)^T$ (Table 6) show that Yin and Tian's stepwise approach and logistic regression perform comparably. In these scenarios, the algorithms achieve a superior performance than when considering simulated data with mean vector $m_1 = (0.2, 0.5, 1.0, 0.7)^T$, as is the case in all tables. Moreover, in this scenario ($m_1 = (0.4, 1.0, 1.5, 1.2)^T$; Table 6), the algorithms discriminate successfully, with the average Youden index achieved by all algorithms being higher than 0.8, with the exception of min-max, which ranges between 0.72 and 0.85. Table 7 shows that these are also scenarios where the combination of biomarkers discriminates satisfactorily. This result could be in line with the literature, where Pinsky and Zhu [42] already unveiled a remarkable increase in performance when considering the combination of highly negatively correlated variables. The results in Table 7 show that the stepwise approaches are worse than the other algorithms, although all of them achieve a perfect or near-perfect performance in some scenarios, with the exception of the min-max approach.

3.1.4. Normal Distributions. Same Means for Diseased and Non-Diseased Population

Table 8 shows the results obtained from the multivariate normal distribution simulations with mean vector $m_1 = (1.0, 1.0, 1.0, 1.0, 1.0)^T$ under the low correlation ($\Sigma_1 = \Sigma_2 = 0.7 \cdot I + 0.3 \cdot J$) and different correlation ($\Sigma_1 = 0.3 \cdot I + 0.7 \cdot J$, $\Sigma_2 = 0.7 \cdot I + 0.3 \cdot J$) scenarios.

Table 8. Normal distributions: Same means: $m_1 = (1.0, 1.0, 1.0, 1.0)^T$.

Size (n_1, n_2)	Mean (SD) Variables	Mean (SD)						Probability Greater than or Equal to Youden Index					
		SLM	SWD	MM	LR	MVN	KS	SLM	SWD	MM	LR	MVN	KS
Same Correlation. Low Correlation ($\Sigma_1 = \Sigma_2 = 0.7 \cdot I + 0.3 \cdot J$)													
(10, 20)	$\bar{x}$ = (0.5178, 0.5176, 0.5180, 0.5150) σ = (0.1439, 0.1440, 0.1377, 0.1427)	**0.8023** (0.0993)	0.7592 (0.1061)	0.704 (0.1177)	0.7158 (0.1335)	0.7128 (0.1256)	0.7505 (0.1165)	**0.5881**	0.1306	0.0969	0.0468	0.0263	0.1113
(30, 30)	$\bar{x}$ = (0.4661, 0.469, 0.4700, 0.4741) σ = (0.1023, 0.1013, 0.1002, 0.0993)	**0.7069** (0.0822)	0.6756 (0.0843)	0.6379 (0.0891)	0.6384 (0.0965)	0.6377 (0.0947)	0.6682 (0.0895)	**0.6635**	0.1258	0.0902	0.0201	0.0167	0.0836
(50, 30)	$\bar{x}$ = (0.4642, 0.4602, 0.4606, 0.4605) σ = (0.0882, 0.0909, 0.0899, 0.0881)	**0.6793** (0.0729)	0.6629 (0.0742)	0.6267 (0.0775)	0.6224 (0.0822)	0.6218 (0.0827)	0.6523 (0.078)	**0.6006**	0.1586	0.1148	0.0118	0.0169	0.0972
(50, 50)	$\bar{x}$ = (0.4491, 0.4483, 0.4448, 0.4479) σ = (0.0787, 0.0795, 0.0778, 0.0782)	**0.6564** (0.0644)	0.634 (0.0674)	0.606 (0.0719)	0.6039 (0.0724)	0.6044 (0.072)	0.6307 (0.0689)	**0.6627**	0.1242	0.1036	<0.01	0.0138	0.0867
(100, 100)	$\bar{x}$ = (0.4266, 0.4258, 0.4270, 0.4252) σ = (0.0536, 0.0578, 0.0580, 0.0581)	**0.6102** (0.0484)	0.5989 (0.0499)	0.5779 (0.0499)	0.5793 (0.0518)	0.5788 (0.0515)	0.5979 (0.0503)	**0.5803**	0.1536	0.1003	0.014	0.0129	0.1389
(500, 500)	$\bar{x}$ = (0.3998, 0.3987, 0.3980, 0.3989) σ = (0.0264, 0.0274, 0.0278, 0.0272)	**0.5556** (0.0247)	0.5536 (0.0251)	0.5387 (0.025)	0.5479 (0.0253)	0.5478 (0.0255)	0.555 (0.0247)	**0.3932**	0.2252	0.0557	0.0337	0.0241	0.2681
Different Correlation ($\Sigma_1 = 0.3 \cdot I + 0.7 \cdot J$, $\Sigma_2 = 0.7 \cdot I + 0.3 \cdot J$)													
(10, 20)	$\bar{x}$ = (0.5149, 0.5141, 0.5172, 0.5203) σ = (0.1438, 0.1394, 0.1453, 0.1453)	**0.7563** (0.1051)	0.7156 (0.1141)	0.7406 (0.1139)	0.6664 (0.1372)	0.6654 (0.1345)	0.7022 (0.1261)	**0.3382**	0.2086	0.2474	0.0590	0.0363	0.1104
(30, 30)	$\bar{x}$ = (0.4640, 0.4658, 0.4703, 0.4718) σ = (0.1016, 0.0991, 0.0979, 0.0981)	0.6641 (0.0853)	0.6353 (0.0877)	**0.6782** (0.0846)	0.5896 (0.0964)	0.5905 (0.0959)	0.6224 (0.0912)	0.3442	0.0746	**0.4400**	0.0210	0.0160	0.1041
(50, 30)	$\bar{x}$ = (0.4625, 0.4576, 0.4586, 0.4598) σ = (0.0902, 0.0924, 0.0862, 0.0906)	0.6424 (0.0735)	0.6256 (0.0752)	**0.6669** (0.0762)	0.5791 (0.0839)	0.5813 (0.0818)	0.6119 (0.0785)	0.2979	0.0744	**0.4954**	0.0119	0.0142	0.1062
(50, 50)	$\bar{x}$ = (0.4481, 0.4442, 0.4439, 0.4457) σ = (0.0805, 0.0794, 0.0770, 0.0822)	0.6145 (0.0653)	0.5936 (0.0686)	**0.6465** (0.0696)	0.5552 (0.0745)	0.5562 (0.075)	0.5826 (0.0728)	0.3163	0.0587	**0.5325**	<0.01	<0.01	0.0800
(100, 100)	$\bar{x}$ = (0.4253, 0.4275, 0.4273, 0.4254) σ = (0.0564, 0.0565, 0.0576, 0.0578)	0.5658 (0.05)	0.5551 (0.0513)	**0.6220** (0.0504)	0.5297 (0.0538)	0.53 (0.0534)	0.5491 (0.0527)	0.1722	0.0516	**0.7520**	<0.01	<0.01	0.0187
(500, 500)	$\bar{x}$ = (0.3994, 0.3992, 0.3994, 0.3993) σ = (0.0275, 0.0270, 0.2720, 0.0267)	0.5085 (0.025)	0.5058 (0.0253)	**0.5870** (0.0240)	0.5004 (0.026)	0.5005 (0.0262)	0.5069 (0.0256)	<0.01	<0.01	**0.9930**	<0.01	<0.01	<0.01

In contrast to the results in Table 2 (different means and low correlation), the results in Table 8 show that the min-max approach performs better in scenarios with biomarkers with the same means. This means that these are scenarios in which the biomarkers have a similar discriminatory capacity, as can be seen in the second column of the table, where the empirical estimates of the Youden index for each biomarker are presented. The table shows that, for scenarios with a low correlation, the min-max algorithm performs similar to the logistic regression and parametric approach under multivariate normality. In this scenario, our proposed stepwise approach dominates the other algorithms. However, this is not the case in the scenario of different covariance matrices ($\Sigma_1 = 0.3 \cdot I + 0.7 \cdot J$, $\Sigma_2 = 0.7 \cdot I + 0.3 \cdot J$), where the min-max approach is the best performer, becoming more and more prominent as the sample size increases. Specifically, in almost 100% of the 1000 simulations of sample size $n_1 = n_2 = 500$, the min-max approach performs best.

3.1.5. Non-Normal Distributions. Different Marginal Distributions

Table 9 shows the results obtained from simulations of multivariate chi-square, normal, gamma and exponential distributions via normal copula with a dependence/correlation parameter between biomarkers of 0.7 and 0.3 for the diseased and non-diseased population, respectively. Biomarkers for the non-diseased population were considered to be marginally distributed as $\chi^2_{0.1}$, $N(0.1, 1)$, $\Gamma(0.1, 1)$ and $Exp(0.1)$, where the considered probability density function for the gamma distribution $X \sim \Gamma(\alpha, \beta)$, using the shape-rate parametrization, is

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1} \exp(-\beta x) \beta^{\alpha}}{\Gamma(\alpha)}, \quad x \geq 0, \alpha, \beta \geq 0 \tag{17}$$

and, for the exponential distribution $X \sim Exp(\lambda)$, λ denoting the rate parameter, the following:

$$f(x; \lambda) = \lambda \exp(-\lambda x), \quad x \geqslant 0, \lambda \geq 0 \tag{18}$$

In the case of the diseased population, two scenarios were considered: $\chi^2_{0.1}$, $N(0.3, 1)$, $\Gamma(0.4, 1)$, $Exp(0.1)$ and $\chi^2_{0.1}$, $N(0.6, 1)$, $\Gamma(0.8, 1)$, $Exp(0.1)$. Since the range of values for each of the four biomarkers was markedly different, it was necessary to normalize the values for each biomarker.

Table 9. Non-normal distributions: Different marginal distributions.

Size (n_1, n_2)	Mean (SD) Variables	Mean (SD)						Probability Greater than or Equal to Youden Index					
		SLM	SWD	MM	LR	MVN	KS	SLM	SWD	MM	LR	MVN	KS
					$N(0.3,1)/\Gamma(0.4,1)$								
(10, 20)	$\bar{x}$ = (0.2111, 0.2710, 0.6032, 0.2119) σ = (0.1249, 0.1339, 0.1215, 0.1247)	**0.7476** (0.099)	0.7004 (0.1076)	0.4910 (0.1258)	0.6047 (0.16)	0.5436 (0.1809)	0.6672 (0.1283)	**0.5544**	0.1723	0.0531	0.0767	0.0187	0.1237
(30, 30)	$\bar{x}$ = (0.1494, 0.2072, 0.5553, 0.1453) σ = (0.0875, 0.0942, 0.0940, 0.0854)	**0.661** (0.0803)	0.6294 (0.0876)	0.4306 (0.0916)	0.5214 (0.1278)	0.4585 (0.1488)	0.6101 (0.1144)	**0.5692**	0.1607	0.0112	0.0367	<0.01	0.2129
(50, 30)	$\bar{x}$ = (0.1364, 0.1939, 0.5476, 0.1328) σ = (0.0757, 0.0828, 0.0876, 0.0745)	**0.6413** (0.0778)	0.6218 (0.0809)	0.4300 (0.0798)	0.5136 (0.1212)	0.4492 (0.1447)	0.6015 (0.1059)	**0.5183**	0.1802	0.0170	0.0312	<0.01	0.2460
(50, 50)	$\bar{x}$ = (0.1161, 0.1752, 0.5384, 0.1126) σ = (0.0656, 0.0720, 0.0744, 0.0610)	**0.6239** (0.0651)	0.6031 (0.0704)	0.3957 (0.0726)	0.4926 (0.1048)	0.4455 (0.1275)	0.5940 (0.0878)	**0.5254**	0.1658	<0.01	0.0142	0.0104	0.2828
(100, 100)	$\bar{x}$ = (0.0847, 0.1466, 0.5216, 0.0829) σ = (0.0459, 0.0551, 0.0537, 0.0460)	**0.5869** (0.047)	0.5755 (0.05)	0.3728 (0.0564)	0.4651 (0.0773)	0.4403 (0.1037)	0.5834 (0.0620)	0.3691	0.1404	<0.01	<0.01	0.0102	**0.4759**
(500, 500)	$\bar{x}$ = (0.0387, 0.1061, 0.4968, 0.0387) σ = (0.0201, 0.0265, 0.0252, 0.0206)	0.5423 (0.0235)	0.5369 (0.0253)	0.3497 (0.0260)	0.4423 (0.0439)	0.4404 (0.0652)	**0.5716** (0.0271)	0.0395	0.0155	<0.01	<0.01	0.0115	**0.9335**
					$N(0.6,1)/\Gamma(0.8,1)$								
(10, 20)	$\bar{x}$ = (0.2111, 0.3666, 0.7690, 0.2119) σ = (0.1249, 0.1412, 0.1005, 0.1247)	**0.8899** (0.0756)	0.8479 (0.0868)	0.5380 (0.1301)	0.804 (0.1413)	0.7612 (0.1563)	0.8296 (0.1097)	**0.5278**	0.1513	<0.01	0.1438	0.0396	0.1282
(30, 30)	$\bar{x}$ = (0.1494, 0.3073, 0.7331, 0.1453) σ = (0.0875, 0.0997, 0.0797, 0.0854)	**0.8406** (0.0685)	0.8013 (0.0723)	0.5206 (0.0909)	0.7626 (0.1083)	0.7249 (0.1202)	0.8042 (0.0807)	**0.6367**	0.1127	<0.01	0.07	0.0154	0.1622
(50, 30)	$\bar{x}$ = (0.1364, 0.2940, 0.7280, 0.1328) σ = (0.0757, 0.0891, 0.0763, 0.0745)	**0.8283** (0.0651)	0.7992 (0.0686)	0.5504 (0.0811)	0.7622 (0.1012)	0.7188 (0.1125)	0.7974 (0.0766)	**0.6465**	0.1148	<0.01	0.0707	0.0111	0.1558
(50, 50)	$\bar{x}$ = (0.1161, 0.2745, 0.7216 ,0.1126) σ = (0.0656, 0.0785, 0.0654, 0.0610)	**0.8178** (0.0554)	0.786 (0.0603)	0.4974 (0.0768)	0.7476 (0.089)	0.7158 (0.0999)	0.7916 (0.0635)	**0.6727**	0.0982	<0.01	0.0583	0.0152	0.1557
(100, 100)	$\bar{x}$ = (0.0847, 0.2516, 0.7089, 0.0829) σ = (0.0459, 0.0605, 0.0457, 0.0460)	**0.7976** (0.0399)	0.771 (0.0424)	0.4836 (0.0579)	0.7354 (0.0651)	0.7114 (0.0733)	0.7805 (0.0433)	**0.7245**	0.0628	<0.01	0.0304	<0.01	0.1727
(500, 500)	$\bar{x}$ = (0.0387, 0.2177, 0.6886, 0.0387) σ = (0.0201, 0.0274, 0.0215, 0.0206)	**0.7774** (0.0193)	0.7547 (0.0217)	0.4635 (0.0256)	0.7221 (0.0347)	0.6929 (0.0373)	0.7698 (0.0194)	**0.8105**	0.0217	<0.01	<0.01	<0.01	0.1645

The results in Table 9 show that our stepwise approach also generally dominates the other approaches in non-normal scenarios with different marginal distributions. It is followed by Yin and Tian's stepwise approach and the non-parametric kernel smoothing approach. Logistic regression outperforms the parametric approach under multivariate normality. The min-max approach is the worst performer.

3.1.6. Non-Normal Distributions. Log-Normal Distributions

Table 10 shows the results obtained from simulated data following a log-normal distribution. Specifically, three scenarios were analyzed under this distribution: independent biomarkers with different means ($\Sigma_1 = \Sigma_2 = I$ and $m_1 = (0.2, 0.5, 1.0, 0.7)^T$), biomarkers correlated with a medium intensity and different means ($\Sigma_1 = \Sigma_2 = 0.5 \cdot I + 0.5 \cdot J$ and $m_1 = (0.2, 0.5, 1.0, 0.7)^T$) and biomarkers correlated with a medium intensity and same means ($\Sigma_1 = \Sigma_2 = 0.5 \cdot I + 0.5 \cdot J$ and $m_1 = (1.0, 1.0, 1.0, 1.0, 1.0)^T$).

The results in Table 10 indicate that, in these scenarios of skewed distributions, our stepwise approach performs significantly better than the other methods. Yin and Tian's stepwise approach performs slightly better than the non-parametric kernel smoothing approach in most scenarios, especially in scenarios where the biomarkers have a similar mean. Logistic regression globally outperforms the parametric approach under multivariate normality and the min-max approach performs better than the logistic approach in biomarker scenarios with the same predictive ability.

From the results provided under sample scenarios of non-normal distributions, it can be deduced that our proposed stepwise approach remains the method that achieves the best overall performance, followed by Yin and Tian's stepwise approach and the non-parametric kernel smoothing approach. The min-max method follows a similar behaviour to that found in normal distribution scenarios, increasing its performance in biomarker samples with a similar predictive ability. Unlike most simulated normal sample data scenarios, in scenarios under non-normal distributions, logistic regression outperforms the parametric approach under multivariate normality.

Table 10. Non-normal distributions: Log-normal distributions.

Size (n_1,n_2)	Mean (SD) Variables	Mean (SD)						Probability Greater than or Equal to Youden Index					
		SLM	SWD	MM	LR	MVN	KS	SLM	SWD	MM	LR	MVN	KS
Different means: $m_1 = (0.2, 0.5, 1.0, 0.7)^T$. Independence ($\Sigma_1 = \Sigma_2 = I$)													
(10, 20)	$\bar{x}$ = (0.2737, 0.3560, 0.5178, 0.4314) σ = (0.1357, 0.1395, 0.1421, 0.1457)	**0.765** (0.1013)	0.7189 (0.1097)	0.627 (0.1222)	0.6548 (0.1448)	0.6284 (0.1524)	0.7051 (0.1194)	**0.6045**	0.1424	0.0867	0.0424	0.0157	0.1082
(30, 30)	$\bar{x}$ = (0.2063, 0.3024, 0.4663, 0.3673) σ = (0.0943, 0.0991, 0.0990, 0.1053)	**0.6545** (0.0835)	0.6235 (0.0857)	0.5527 (0.0959)	0.5651 (0.1014)	0.5422 (0.1088)	0.6162 (0.0892)	**0.6476**	0.1413	0.0772	0.0152	<0.01	0.1121
(50, 30)	$\bar{x}$ = (0.1933, 0.2975, 0.4630, 0.3603) σ = (0.0846, 0.0937, 0.0894, 0.0898)	**0.6289** (0.0742)	0.6137 (0.0758)	0.5397 (0.0816)	0.5539 (0.0888)	0.5313 (0.0942)	0.6023 (0.0798)	**0.5810**	0.1785	0.0862	0.0164	<0.01	0.1317
(50, 50)	$\bar{x}$ = (0.1764, 0.2784, 0.4484, 0.3458) σ = (0.0736, 0.0789, 0.0774, 0.0796)	**0.605** (0.0663)	0.5829 (0.0682)	0.5163 (0.0737)	0.5308 (0.0796)	0.5139 (0.0832)	0.5735 (0.0715)	**0.6739**	0.1518	0.0638	<0.01	<0.01	0.0996
(100, 100)	$\bar{x}$ = (0.1487, 0.2498, 0.4254, 0.3214) σ = (0.0533, 0.0590, 0.0560, 0.0590)	**0.5507** (0.0498)	0.5399 (0.0514)	0.4802 (0.0528)	0.5027 (0.0560)	0.4911 (0.0578)	0.5322 (0.0526)	**0.6514**	0.1676	0.0418	0.0105	<0.01	0.1223
(500, 500)	$\bar{x}$ = (0.1062, 0.2185, 0.3991, 0.2925) σ = (0.0257, 0.0282, 0.0274, 0.0280)	**0.4926** (0.0255)	0.4904 (0.0255)	0.4412 (0.0267)	0.4779 (0.0265)	0.4745 (0.0269)	0.4890 (0.0259)	**0.5157**	0.2015	<0.01	0.0257	0.0198	0.2308
Different means: $m_1 = (0.2, 0.5, 1.0, 0.7)^T$. Medium Correlation ($\Sigma_1 = \Sigma_2 = 0.5 \cdot I + 0.5 \cdot J$)													
(10, 20)	$\bar{x}$ = (0.2713, 0.3657, 0.5268, 0.4280) σ = (0.1353, 0.1477, 0.1423, 0.1427)	**0.7319** (0.1064)	0.6647 (0.1185)	0.5354 (0.1314)	0.6227 (0.1429)	0.5784 (0.1568)	0.6732 (0.1191)	**0.5570**	0.1647	0.0280	0.0741	0.0245	0.1518
(30, 30)	$\bar{x}$ = (0.2032, 0.3020, 0.4665, 0.3668) σ = (0.0933, 0.1014, 0.0991, 0.1008)	**0.6177** (0.0836)	0.5641 (0.0885)	0.4490 (0.0981)	0.5177 (0.1016)	0.4834 (0.1153)	0.5712 (0.0899)	**0.5712**	0.1799	0.0124	0.0354	0.0101	0.1911
(50, 30)	$\bar{x}$ = (0.1931, 0.2928, 0.4581, 0.3632) σ = (0.0847, 0.0908, 0.0883, 0.0905)	**0.5848** (0.0767)	0.5579 (0.0786)	0.4333 (0.0838)	0.5066 (0.0896)	0.4713 (0.1036)	0.5564 (0.082)	**0.5245**	0.2395	<0.01	0.0220	0.011	0.1972
(50, 50)	$\bar{x}$ = (0.1753, 0.2761, 0.4472, 0.3495) σ = (0.0742, 0.0798, 0.0787, 0.0791)	**0.5619** (0.0676)	0.5243 (0.0712)	0.4121 (0.0767)	0.4864 (0.0792)	0.4546 (0.0899)	0.5268 (0.0745)	**0.5553**	0.1843	<0.01	0.0260	<0.01	0.2243
(100, 100)	$\bar{x}$ = (0.1449, 0.2487, 0.4250, 0.3236) σ = (0.0520, 0.0578, 0.0591, 0.0595)	**0.5089** (0.0539)	0.4845 (0.056)	0.378 (0.0567)	0.4567 (0.0608)	0.4345 (0.0677)	0.4859 (0.0582)	**0.5573**	0.2003	<0.01	0.0208	<0.01	0.2118
(500, 500)	$\bar{x}$ = (0.1063, 0.2187, 0.3994, 0.2921) σ = (0.0261, 0.0280, 0.0264, 0.0279)	**0.4446** (0.0261)	0.4374 (0.0257)	0.3329 (0.0273)	0.4285 (0.0265)	0.4202 (0.0282)	0.4394 (0.0265)	**0.5233**	0.1825	<0.01	0.0315	<0.01	0.2538
Same means: $m_1 = (1.0, 1.0, 1.0, 1.0)^T$. Medium Correlation ($\Sigma_1 = \Sigma_2 = 0.5 \cdot I + 0.5 \cdot J$)													
(10, 20)	$\bar{x}$ = (0.5224, 0.5172, 0.5268, 0.5197) σ = (0.1405, 0.1442, 0.1423, 0.1404)	**0.7619** (0.1032)	0.7249 (0.1127)	0.66 (0.1267)	0.663 (0.1402)	0.6254 (0.1508)	0.7041 (0.1187)	**0.5180**	0.2036	0.0794	0.0675	0.0183	0.1132
(30, 30)	$\bar{x}$ = (0.4685, 0.4669, 0.4665, 0.4640) σ = (0.1024, 0.1025, 0.0991, 0.0991)	**0.6624** (0.0851)	0.6291 (0.0878)	0.5882 (0.0926)	0.5676 (0.101)	0.5385 (0.1092)	0.6089 (0.0939)	**0.5416**	0.2341	0.0776	0.0251	<0.01	0.1123
(50, 30)	$\bar{x}$ = (0.4586, 0.4621, 0.4581, 0.4627) σ = (0.0878, 0.0908, 0.0883, 0.0878)	**0.6337** (0.0759)	0.6167 (0.0773)	0.5758 (0.082)	0.5566 (0.0871)	0.5263 (0.102)	0.5949 (0.083)	**0.5167**	0.2714	0.0851	0.0171	0.0107	0.0991
(50, 50)	$\bar{x}$ = (0.4459, 0.4477, 0.4472, 0.4479) σ = (0.0769, 0.0810, 0.0787, 0.0789)	**0.6089** (0.0673)	0.5887 (0.0715)	0.5562 (0.0737)	0.5335 (0.0795)	0.5061 (0.087)	0.5688 (0.0758)	**0.5359**	0.2648	0.0887	0.0127	<0.01	0.0929
(100, 100)	$\bar{x}$ = (0.4243, 0.4235, 0.4250, 0.4263) σ = (0.0569, 0.0572, 0.0591, 0.0597)	**0.5598** (0.0498)	0.5457 (0.0525)	0.5266 (0.0537)	0.5081 (0.0565)	0.4879 (0.0638)	0.5331 (0.0544)	**0.5434**	0.2628	0.0996	0.0128	<0.01	0.0772
(500, 500)	$\bar{x}$ = (0.3993, 0.3990, 0.3994, 0.3992) σ = (0.0268, 0.0277, 0.0264, 0.0270)	**0.4963** (0.0254)	0.4934 (0.0255)	0.4883 (0.0256)	0.4819 (0.0265)	0.4748 (0.0277)	0.4908 (0.0259)	**0.4802**	0.2588	0.1300	0.0135	<0.01	0.1105

3.2. Simulations. Validation

Tables 11–16 show the results of the validation of each algorithm for every simulated data scenario. In particular, for all simulated setting of parameters, using 100 random samples, and for each method, the mean and standard deviation (in brackets) of the maximum Youden indexes obtained in the analysis of the 100 validation samples are presented.

These results and conclusions drawn in terms of validation in the simulated scenarios are presented below.

3.2.1. Normal Distributions. Different Means and Equal Positive Correlations for Diseased and Non-Diseased Population

The results in Table 11 show that, for normal simulated data, different means and equal positive correlation, the logistic regression and the parametric approach under multivariate normality outperform the rest of the methods in all scenarios for small or large sample sizes. The non-parametric kernel smoothing approach shows lower but comparable results to the logistic regression and non-parametric approach, especially in large sample sizes. Our stepwise approach outperforms Yin and Tian's stepwise approach, especially and significantly in the high correlation scenario. Our stepwise approach is closer in performance to the non-parametric kernel smoothing approach for large sample sizes. The min-max approach is the one with the worst results in such scenarios.

Table 11. Normal distributions: Different means and equal positive correlations. Validation.

Size (n_1, n_2)	**Independence ($\Sigma_1 = \Sigma_2 = I$)**					
	SLM	**SWD**	**MM**	**LR**	**MVN**	**KS**
	$m_1 = (0.2, 0.5, 1.0, 0.7)^T$					
(50, 50)	0.4270 (0.0951)	0.4206 (0.1134)	0.3696 (0.1037)	**0.4530** (0.0919)	0.4520 (0.0934)	0.4434 (0.0911)
(500, 500)	0.4823 (0.0304)	0.4804 (0.0291)	0.4103 (0.0308)	0.4882 (0.0282)	**0.4895** (0.0297)	0.4863 (0.0301)
	$m_1 = (0.4, 1.0, 1.5, 1.2)^T$					
(50, 50)	0.6610 (0.0842)	0.6610 (0.0880)	0.6024 (0.0908)	0.6990 (0.0779)	**0.6994** (0.0787)	0.6902 (0.0773)
(500, 500)	0.7163 (0.0224)	0.7159 (0.0228)	0.6418 (0.0253)	0.7238 (0.0205)	**0.7249** (0.0200)	0.7223 (0.0207)
	Low correlation ($\Sigma_1 = \Sigma_2 = 0.7 \cdot I + 0.3 \cdot J$)					
	SLM	SWD	MM	LR	MVN	KS
	$m_1 = (0.2, 0.5, 1.0, 0.7)^T$					
(50, 50)	0.3430 (0.0969)	0.3376 (0.1041)	0.2642 (0.1211)	0.3686 (0.0949)	**0.3736** (0.0924)	0.3586 (0.1012)
(500, 500)	0.4074 (0.0285)	0.4056 (0.0303)	0.3189 (0.0322)	0.4149 (0.0309)	**0.4155** (0.0314)	0.4131 (0.0292)
	$m_1 = (0.4, 1.0, 1.5, 1.2)^T$					
(50, 50)	0.5576 (0.0939)	0.5510 (0.1024)	0.5056 (0.0979)	**0.5790** (0.0921)	**0.5790** (0.0853)	0.5740 (0.0955)
(500, 500)	0.6084 (0.0294)	0.6079 (0.0284)	0.5213 (0.0289)	**0.6183** (0.0286)	0.6181 (0.0284)	0.6161 (0.0289)
	Medium correlation ($\Sigma_1 = \Sigma_2 = 0.5 \cdot I + 0.5 \cdot J$)					
	SLM	SWD	MM	LR	MVN	KS
	$m_1 = (0.2, 0.5, 1.0, 0.7)^T$					
(50, 50)	0.3618 (0.0984)	0.3442 (0.1009)	0.2474 (0.1072)	**0.3750** (0.0854)	0.3678 (0.0839)	0.3640 (0.0880)
(500, 500)	0.4088 (0.0351)	0.4039 (0.0327)	0.2924 (0.0323)	0.4178 (0.0334)	**0.4189** (0.0340)	0.4154 (0.0336)
	$m_1 = (0.4, 1.0, 1.5, 1.2)^T$					
(50, 50)	0.5490 (0.0921)	0.5340 (0.1002)	0.4584 (0.0994)	0.5660 (0.0830)	**0.5684** (0.0864)	0.5584 (0.0876)
(500, 500)	0.5936 (0.0303)	0.5914 (0.0303)	0.4840 (0.0293)	**0.6063** (0.0319)	0.6059 (0.0301)	0.6034 (0.0286)
	High correlation ($\Sigma_1 = \Sigma_2 = 0.3 \cdot I + 0.7 \cdot J$)					
	SLM	SWD	MM	LR	MVN	KS
	$m_1 = (0.2, 0.5, 1.0, 0.7)^T$					
(50, 50)	0.4150 (0.0946)	0.3430 (0.1203)	0.2522 (0.1081)	0.4296 (0.0952)	**0.4310**(0.0989)	0.4136 (0.0962)
(500, 500)	0.4588 (0.0307)	0.4234 (0.0386)	0.2899 (0.0317)	0.4662 (0.0260)	**0.4666** (0.0258)	0.4632 (0.0280)
	$m_1 = (0.4, 1.0, 1.5, 1.2)^T$					
(50, 50)	0.5892 (0.09036)	0.5364 (0.1098)	0.4542 (0.0957)	**0.6218** (0.0839)	0.6196 (0.0851)	0.6134 (0.0892)
(500, 500)	0.6285 (0.0234)	0.6171 (0.0307)	0.4823 (0.0315)	0.6456 (0.0220)	**0.6461** (0.0213)	0.6460 (0.0227)

3.2.2. Normal Distributions. Different Means and Unequal Positive Correlations for Diseased and Non-Diseased Population

For normal simulated data, different means and unequal positive correlation for diseased and non-diseased population, the results displayed in Table 12 show logistic regression and the parametric approach under multivariate normality as the best models, and as very similar to the non-parametric kernel smoothing approach, with the stepwise approaches slightly worse and the min-max method with clearly lower values.

Table 12. Normal distributions: Different means and unequal positive correlations. Validation.

Size (n_1, n_2)	**Different Correlation ($\Sigma_1 = 0.3 \cdot I + 0.7 \cdot J, \Sigma_2 = 0.7 \cdot I + 0.3 \cdot J$)**					
	SLM	**SWD**	**MM**	**LR**	**MVN**	**KS**
	$m_1 = (0.2, 0.5, 1.0, 0.7)^T$					
(50, 50)	0.3528 (0.1106)	0.3532 (0.1072)	0.3236 (0.1012)	**0.3878** (0.1079)	0.3814 (0.1061)	0.3832 (0.1127)
(500, 500)	0.4162 (0.0278)	0.4128 (0.0318)	0.3534 (0.0329)	0.4224 (0.0282)	**0.4233** (0.0295)	0.4120 (0.0281)
	$m_1 = (0.4, 1.0, 1.5, 1.2)^T$					
(50, 50)	0.5500 (0.1030)	0.5218 (0.1041)	0.4328 (0.0880)	**0.5792** (0.1014)	0.5730 (0.1012)	0.5700 (0.1025)
(500, 500)	0.5974 (0.0295)	0.5966 (0.0299)	0.4757 (0.0320)	**0.6062** (0.0276)	0.6053 (0.0275)	0.6038 (0.0271)

3.2.3. Normal Distributions. Different Means and Equal Negative Correlations for Diseased and Non-Diseased Population

The performance of the algorithms for normal simulated data, different means and equal negative correlation was very similar to previous results. It can be seen in Table 13 that logistic regression and the parametric approach under multivariate normality were the best models, followed by the non-parametric kernel smoothing approach, our stepwise approach and Yin and Tian's stepwise approach, with worse results for the min-max algorithm.

Table 13. Normal distributions: Different means and equal negative correlations. Validation.

Size (n_1, n_2)	**Negative Correlation (-0.1)**					
	SLM	**SWD**	**MM**	**LR**	**MVN**	**KS**
	$m_1 = (0.2, 0.5, 1.0, 0.7)^T$					
(50, 50)	0.4822 (0.0990)	0.4670 (0.1028)	0.4316 (0.0939)	**0.5168** (0.0944)	0.5160 (0.0920)	0.5010 (0.0920)
(500, 500)	0.5457 (0.0263)	0.5452 (0.0268)	0.4631 (0.0296)	0.5545 (0.0255)	**0.5558** (0.0264)	0.5504 (0.0275)
	$m_1 = (0.4, 1.0, 1.5, 1.2)^T$					
(50, 50)	0.7444 (0.0800)	0.7388 (0.0725)	0.6788 (0.0874)	0.7694 (0.0671)	**0.7730** (0.0653)	0.7702 (0.0611)
(500, 500)	0.7960 (0.0209)	0.7953 (0.0213)	0.7085 (0.0265)	**0.8015** (0.0204)	**0.8015** (0.0216)	0.7990 (0.0194)
	Negative correlation (-0.3)					
	SLM	SWD	MM	LR	MVN	KS
	$m_1 = (0.2, 0.5, 1.0, 0.7)^T$					
(50, 50)	0.8696 (0.0790)	0.8046 (0.1118)	0.6602 (0.0787)	0.9210 (0.0426)	**0.9284** (0.0374)	0.9198 (0.0444)
(500,500)	0.9192 (0.0242)	0.9107 (0.0278)	0.6930 (0.0239)	**0.9424** (0.0110)	0.9423 (0.0117)	0.9417 (0.0114)
	$m_1 = (0.4, 1.0, 1.5, 1.2)^T$					
(50, 50)	0.9382 (0.0600)	0.9462 (0.0485)	0.8754 (0.0545)	0.9544 (0.0410)	**0.9734** (0.0338)	0.9646 (0.0443)
(500,500)	0.9950 (0.0043)	0.9943 (0.0047)	0.9001 (0.0153)	0.9948 (0.0049)	**0.9975** (0.0030)	0.9958 (0.0047)

3.2.4. Normal Distributions. Same Means for Diseased and Non-Diseased Population

Regarding scenarios with normal simulated data and same means for the diseased and non-diseased populations, the results in Table 14 present clear differences with previous simulations. For scenarios of the same correlation, the min-max algorithm is not the worst and all algorithms show a very similar mean Youden index in large samples. However, for scenarios of different correlations, the min-max algorithm clearly outperforms the rest of the algorithms.

Table 14. Normal distributions: Same means. Validation.

Size (n_1, n_2)	**Same Means ($m_1 = (1.0, 1.0, 1.0, 1.0)^T$)**					
	SLM	**SWD**	**MM**	**LR**	**MVN**	**KS**
	Same Correlation. Low Correlation ($\Sigma_1 = \Sigma_2 = 0.7 \cdot I + 0.3 \cdot J$)					
(50, 50)	0.4626 (0.1062)	0.4376 (0.1081)	0.4794 (0.0954)	0.4878 (0.0900)	**0.4882** (0.0944)	0.4852 (0.0951)
(500 ,500)	0.5129 (0.0305)	0.5174 (0.0318)	0.5073 (0.0278)	**0.5254** (0.0285)	**0.5254** (0.0284)	0.5219 (0.0283)
	Different Correlation ($\Sigma_1 = 0.3 \cdot I + 0.7 \cdot J, \Sigma_2 = 0.7 \cdot I + 0.3 \cdot J$)					
(50, 50)	0.4030 (0.1110)	0.4102 (0.1057)	**0.5220** (0.0966)	0.4340 (0.1029)	0.4402 (0.1027)	0.4312 (0.1018)
(500, 500)	0.4726 (0.0281)	0.4697 (0.0280)	**0.5609** (0.0285)	0.4783 (0.0259)	0.4793 (0.0260)	0.4753 (0.0256)

Thus, in summary, for normal simulated data, our stepwise approach, Yin and Tian's stepwise model, logistic regression, parametric under multivariate normality and non-parametric kernel smoothing algorithms showed a close performance, with the best results for logistic regression and the parametric approach under multivariate normality, an intermediate position for the kernel smoothing algorithm and lower values for our proposed stepwise approach, which is still better than Yin and Tian's stepwise algorithm in most cases,

and significantly for high correlations. By contrast, the min-max algorithm has a worse performance for scenarios with different means, but is clearly superior for simulations generated with the same mean for disease markers and different correlations for disease and non-disease populations.

3.2.5. Non-Normal Distributions. Different Marginal Distributions

Table 15 shows results for the non-normal scenario with different marginal distributions, which is a scenario that is probably closer to the reality of the actual data, where asymmetries occur when patients have different degrees of a disease. In this cases, the stepwise approaches clearly outperform the rest of the algorithms, with the better results for our proposed stepwise algorithm. For large sample sizes, the non-parametric kernel smoothing shows markedly superior results to the logistic and the parametric approach. As in some previous cases, the min-max method fails to provide similar results in these scenarios.

Table 15. Non-normal distributions: Different marginal distributions. Validation.

Size (n_1, n_2)	Different Marginal Distributions					
	SLM	**SWD**	**MM**	**LR**	**MVN**	**KS**
	$N(0.3,1)/\Gamma(0.4,1)$					
(50, 50)	**0.4732 (0.0860)**	0.4684 (0.0891)	0.2374 (0.1158)	0.3656 (0.1243)	0.3316 (0.1534)	0.3146 (0.2140)
(500, 500)	**0.5095 (0.0277)**	0.5018 (0.0297)	0.3180 (0.0442)	0.4137 (0.0459)	0.4285 (0.0671)	0.4787 (0.1191)
	$N(0.6,1)/\Gamma(0.8,1)$					
(50, 50)	**0.7058 (0.0848)**	0.6794 (0.0877)	0.3716 (0.1194)	0.6572 (0.1080)	0.6368 (0.1079)	0.6716 (0.1207)
(500, 500)	**0.7568 (0.0231)**	0.7351 (0.0229)	0.4350 (0.04530)	0.7065 (0.0363)	0.6807 (0.0360)	0.7469 (0.0304)

3.2.6. Non-Normal Distributions. Log-Normal Distributions

Table 16 shows the results for the simulated log-normal distributions. Similar conclusions can be drawn for the simulated normal data. We can infer that, for distributions that can be converted into normal distributions by means of monotonic transformations, we expect to find similar conclusions as for the simulations under the normality hypothesis.

Table 16. Non-normal distributions: Log-normal distributions. Validation.

Size (n_1, n_2)	Log-Normal Distributions					
	SLM	**SWD**	**MM**	**LR**	**MVN**	**KS**
	Different means: $m_1 = (0.2, 0.5, 1.0, 0.7)^T$. Independence ($\Sigma_1 = \Sigma_2 = I$)					
(50, 50)	0.4022 (0.1019)	0.4078 (0.1041)	0.3658 (0.1060)	**0.4112 (0.0936)**	0.4034 (0.0903)	0.3914 (0.1142)
(500, 500)	0.4504 (0.0296)	0.4506 (0.0315)	0.4096 (0.0324)	**0.4562 (0.0300)**	0.4541 (0.0321)	0.4507 (0.0534)
	Different means: $m_1 = (0.2, 0.5, 1.0, 0.7)^T$. Medium correlation ($\Sigma_1 = \Sigma_2 = 0.5 \cdot I + 0.5 \cdot J$)					
(50, 50)	0.3460 (0.0103)	0.3454 (0.1056)	0.2574 (0.1024017)	**0.3482 (0.1065)**	0.3370 (0.1100)	0.3374 (0.1133)
(500, 500)	0.3990 (0.0345)	0.3954 (0.0358)	0.2924 (0.0336)	0.3990 (0.0372)	0.3960 (0.0367)	**0.4014 (0.0348)**
	Same means: $m_1 = (1.0, 1.0, 1.0, 1.0)^T$. Medium correlation ($\Sigma_1 = \Sigma_2 = 0.5 \cdot I + 0.5 \cdot J$)					
(50, 50)	0.3890 (0.1035)	0.3756 (0.1029)	**0.4170 (0.1046)**	0.4102 (0.0969)	0.3796 (0.1187)	0.3584 (0.1133)
(500, 500)	0.4465 (0.0306)	0.4515 (0.0315)	0.4548 (0.0282)	**0.4570 (0.0309)**	0.4514 (0.0317)	0.4545 (0.0317)

3.3. Computational Times

In addition to the performance of the algorithms in terms of the Youden index, in practice, it can be important to also consider the computational time taken by the algorithm to be used. Although our proposal is an algorithm of an extensive search, when k is the number of values of β_i to be considered and p is the number of markers, the number of Youden indexes necessary to estimate the parameters of the model is reduced from the $p \cdot k^{p-1}$ order using the comprehensive Pepe and Thompson algorithm [13] to the $k \cdot (p-1)(\frac{3}{2}p - 1)$ order using the stepwise procedure.

Without a loss of generality, Table 17 shows the average computational time of 1000 simulations of each of the analyzed algorithms for the scenario of normal distributions, a low positive correlation and vector of means ($m_1 = (0.2, 0.5, 1.0, 0.7)$) (Table 2), both for the smallest sample size ($n_1 = 10, n_2 = 20$) and for the largest sample size ($n_1 = 500, n_2 = 500$).

Table 17. Total computational time for each algorithm (estimated by mean of 1000 samples).

	Computational Times (min)					
	SLM	**SWD**	**MM**	**LR**	**MVN**	**KS**
$(n_1, n_2) = (10, 20)$	17.1157	0.096	0.014	0.00003	0.00004	0.0003
$(n_1, n_2) = (500, 500)$	0.5939	0.096	0.033	0.00004	0.00004	0.0007

Table 17 shows that the stepwise algorithms have a longer computational time than the other algorithms. Our proposed algorithm entails a noticeably higher computational time compared to Yin and Tian's stepwise approach. This difference in the computational time is due to the biomarker search that optimizes the linear combination at each step and the handling of ties in our algorithm, which gets worse at small sample sizes where ties are more common. As a consequence, there is a high disparity between computational times with small sample sizes. This computational burden increases significantly for a larger number of biomarkers. However, although this high computational time presents a limitation in our algorithm, it addresses a correct handling of ties, leading to better discriminatory ability results. Furthermore, the computational time of a single simulation is, for four biomarkers, in any case, addressable. It should also be noted that the computational times of derivative-based numerical search methods (such as the non-parametric kernel smoothing approach) significantly depend on the initial values.

3.4. Application in Clinical Diagnosis Cases

Figures 1 and 2 show the distribution of each biomarker for the Duchenne muscular dystrophy and prostate cancer dataset, respectively, where disease refers to clinically significant prostate cancer.

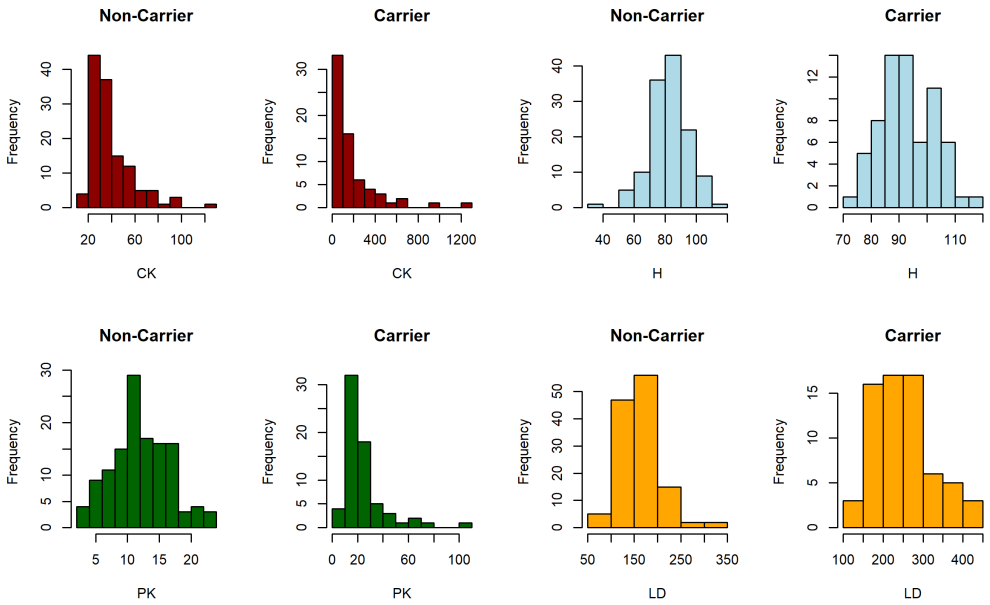


Figure 1. Marginal distributions of biomarkers. DMD dataset. CK: serum creatine kinase, H: haemopexin, PK: pyruvate kinase, LD: lactate dehydrogenase.

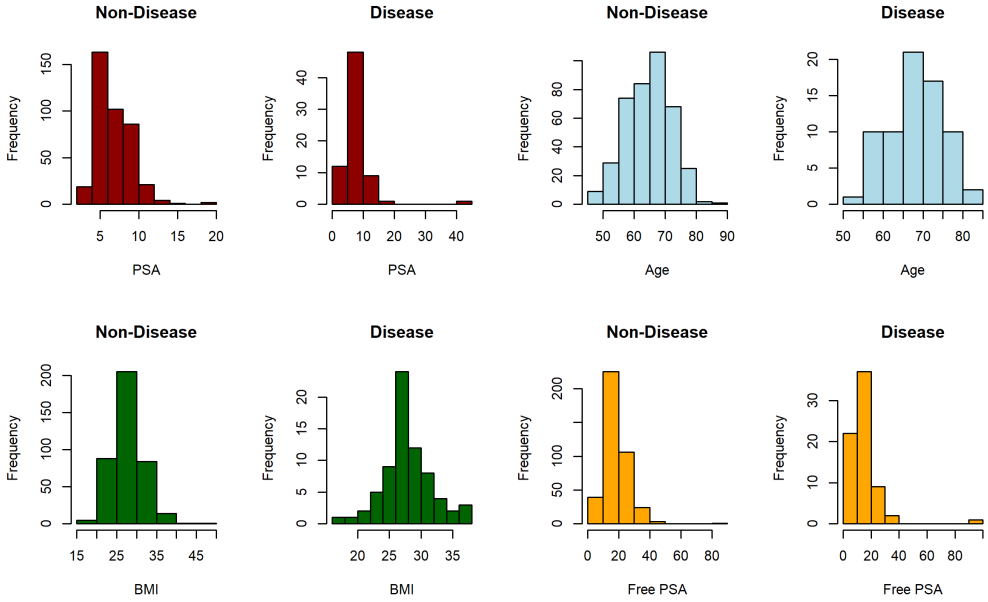


Figure 2. Marginal distributions of biomarkers. Prostate cancer dataset. PSA: prostate specific antigen, Age: age in years, BMI: body mass index, Free PSA: percentage of Free PSA.

Tables 18 and 19 show the empirical estimates of the Youden index of each biomarker and the optimal cut-off point (threshold), as well as the characteristics of each of them in terms of mean, standard deviations (SD) and correlations between them for both the disease and non-disease group, considered as a non-carrier for the Duchenne muscular dystrophy dataset and non-clinically significant prostate cancer for the prostate cancer dataset, respectively.

Table 18. DMD dataset information.

			Non-Carrier		**Carrier**	
	Youden	*Threshold*	*Mean*	*SD*	*Mean*	*SD*
CK	0.6124	57	36.6102	18.6006	185.791	226.9330
H	0.4172	87.5	82.3072	12.2403	92.9303	9.8576
PK	0.5079	16.7	12.1447	4.3935	23.9310	17.2122
LD	0.5776	188	164.5748	41.3686	250.9403	72.4368
			Correlations			
Non-Carrier	r_{CK-H}	r_{CK-PK}	r_{CK-LD}	r_{H-PK}	r_{H-LD}	r_{PK-LD}
	-0.3340	0.1029	0.1987	0.0812	0.1824	0.2188
Carrier	r_{CK-H}	r_{CK-PK}	r_{CK-LD}	r_{H-PK}	r_{H-LD}	r_{PK-LD}
	-0.1364	0.6953	0.4851	-0.118	-0.1048	0.4813

r_{CK-H}, r_{CK-PK}, r_{CK-LD}, r_{H-PK}, r_{H-LD}, r_{PK-LD} denote the correlations between *CK* and *H* biomarkers, *CK* and *PK* biomarkers, *CK* and *LD* biomarkers, *H* and *PK* biomarkers, *H* and *LD* biomarkers and *PK* and *LD* biomarkers, respectively.

Table 19. Prostate cancer dataset information.

			Non-Cancer		Cancer	
	Youden	*Threshold*	*Mean*	*SD*	*Mean*	*SD*
PSA	0.1571	9.45	6.7875	2.3160	7.9887	5.2761
Age	0.2202	68	65.0804	7.2840	68.8732	6.6846
BMI	0.0953	25.83	27.8590	3.8243	27.7559	3.8756
Free PSA	0.4007	13.95	18.3629	7.5917	14.7190	11.4067
			Correlations			
Non-Cancer	$r_{PSA-Age}$	$r_{PSA-BMI}$	$r_{PSA-FreePSA}$	$r_{Age-BMI}$	$r_{Age-FreePSA}$	$r_{BMI-FreePSA}$
	0.0901	−0.1179	−0.1127	0.0536	0.0894	0.0694
Cancer	$r_{PSA-Age}$	$r_{PSA-BMI}$	$r_{PSA-FreePSA}$	$r_{Age-BMI}$	$r_{Age-FreePSA}$	$r_{BMI-FreePSA}$
	−0.1985	0.0896	−0.0756	0.0758	0.2478	−0.0767

$r_{PSA-Age}, r_{PSA-BMI}, r_{PSA-FreePSA}, r_{Age-BMI}, r_{Age-FreePSA}, r_{BMI-FreePSA}$ denote the correlations between PSA and Age biomarkers, PSA and BMI biomarkers, PSA and $FreePSA$ biomarkers, Age and BMI biomarkers, Age and $FreePSA$ biomarkers and BMI and $FreePSA$ biomarkers, respectively.

Concerning the performance achieved by each method, Tables 20 and 21 present the linear combination for the optimal cut-off point that maximizes the Youden index, as well as the sensitivity and specificity values achieved, for the Duchenne muscular dystrophy and prostate cancer dataset, respectively. Tables 22 and 23 show these metrics after applying the 10-fold cross validation procedure.

Table 20. Linear combination that maximizes the Youden index for each method. DMD dataset.

	Optimal Linear Combination	Youden	Sensitivity	Specificity
SLM	$0.57 \times CK + H + 0.65 \times PK + 0.08 \times LD$	0.8255	0.8806	0.9449
SWD	$0.36 \times CK + 0.82 \times H + PK + 0.1296 \times LD$	0.8184	0.8657	0.9528
MM	$0.14 \times \max\{\overline{CK}, \overline{H}, \overline{PK}, \overline{LD}\} + \min\{\overline{CK}, \overline{H}, \overline{PK}, \overline{LD}\}$	0.7335	0.8358	0.8976
LR	$0.0482 \times CK + 0.1039 \times H + 0.0992 \times PK + 0.0138 \times LD$	0.8106	0.8657	0.9449
MVN	$0.0956 \times CK + 0.126 \times H + 0.1847 \times PK + 0.0316 \times LD$	0.7878	0.8507	0.9370
KS	$1.1699 \times CK + 3.1787 \times H + 3.819 \times PK + 0.5899 \times LD$	0.8035	0.8507	0.9528

$\overline{CK}, \overline{H}, \overline{PK}, \overline{LD}$: biomarkers normalized after min-max scaling. The results rounded to four decimal places are displayed.

Table 21. Linear combination that maximizes the Youden index for each method. Prostate cancer dataset.

	Optimal Linear Combination	Youden	Sensitivity	Specificity
SLM	$0.04 \times PSA + 0.48 \times Age - 0.01 \times BMI - FreePSA$	0.4857	0.7746	0.7111
SWD	$PSA + 0.84 \times Age + 0.07 \times BMI - FreePSA$	0.4319	0.7887	0.6432
MM	$\max\{\overline{PSA}, \overline{Age}, \overline{BMI}, \overline{FreePSA}\} - \min\{\overline{PSA}, \overline{Age}, \overline{BMI}, \overline{FreePSA}\}$	0.2986	0.5775	0.7211
LR	$0.0881 \times PSA + 0.0803 \times Age - 0.0079 \times BMI - 0.0755 \times FreePSA$	0.4284	0.7324	0.6960
MVN	$0.2605 \times PSA + 0.335 \times Age - 0.086 \times BMI - 0.162 \times FreePSA$	0.3660	0.6901	0.6759
KS	$1.1737 \times PSA + 27.7107 \times Age - 7.7898 \times BMI - 51.2465 \times FreePSA$	0.4681	0.7746	0.6935

$\overline{PSA}, \overline{Age}, \overline{BMI}, \overline{FreePSA}$: biomarkers normalized after min-max scaling. The results rounded to four decimal places are displayed.

Table 22. Ten-fold cross validation. DMD dataset.

	10-Fold Cross Validation. DMD Dataset.		
	Youden	**Sensitivity**	**Specificity**
SLM	0.7611	0.8576	0.9135
SWD	0.7301	0.8167	0.9135
MM	0.6215	0.7786	0.8429
LR	**0.7861**	0.8476	0.9345
MVN	0.7391	0.8167	0.9224
KS	0.7635	0.8333	0.9301

Table 23. Ten-fold cross validation. Prostate dataset.

	10-Fold Cross Validation. Prostate Dataset.		
	Youden	**Sensitivity**	**Specificity**
SLM	**0.3844**	0.6786	0.7058
SWD	0.3628	0.6946	0.6681
MM	0.2247	0.4661	0.7586
LR	0.3327	0.6768	0.6559
MVN	0.2785	0.6625	0.6160
KS	0.3820	0.6911	0.6910

3.4.1. Duchenne Muscular Dystrophy Dataset

The mean values of each biomarker and the correlations between them are very different and differ between carriers and non-carriers. Likewise, the variances of the four biomarkers are very different and, therefore, it is necessary to normalize the values of each variable before applying the min-max method. In this way, different units of measurement are avoided so that a correct use of the min-max algorithm is made, where all biomarkers must be in the same unit. The estimates of the Youden index of each biomarker (CK, H, PK, LD) in a univariate way were 0.6124, 0.4172, 0.5079 and 0.5776.

The linear methods (combination of biomarkers) achieved a remarkable Youden index in training data, with values above 0.8 for most of them. Stepwise methods followed by logistic regression and the non-parametric method based on kernel smoothing are the ones that obtained the best results. Our proposed stepwise approach achieved the best performance in training data (Youden index = 0.8255) with the linear combination $0.57 \times CK + H + 0.65 \times PK + 0.08 \times LD$, but the logistic regression showed the best result in a 10-fold cross validation procedure (Youden index = 0.7861) with the linear combination $0.0482 \times CK + 0.1039 \times H + 0.0992 \times PK + 0.0138 \times LD$. It is followed by the kernel algorithm and our stepwise approach. These results are in concordance with those of normal simulated data or variables that can be converted into normal distributions by means of monotonic transformations.

3.4.2. Prostate Cancer Dataset

Although to a lesser extent than the previous example, there is a notable difference between the variances of the biomarkers, so the values were also normalized before applying the min-max approach. The correlations between biomarkers are close to zero (independent biomarkers). The Youden index estimates for each biomarker (PSA, age, BMI, free PSA) in a univariate way were lower than the previous data set: 0.1571, 0.2202, 0.0953 and 0.0141.

Our proposed stepwise algorithm and the non-parametric method based on kernel smoothing dominate all of the other methods. Our algorithm achieved the best performance in training and validation data (maximum Youden index = 0.4857, 0.3844 for training and validation data respectively) with the linear combination $0.04 \times PSA + 0.48 \times Age - 0.01 \times BMI - FreePSA$. In these cases, PSA and free PSA are markers that usually present a marked asymmetry, showing a greater or lower degree of progress of the cancer disease; in this scenario, simulated data also showed the superiority of the stepwise algorithm.

4. Discussion

Although continuous markers usually provide better adjusted predictions, in classification problems, the ultimate goal is to assign a class 0/1 for any individual. Choosing threshold probabilities to dichotomize a predictive or prognostic model is the key to solve this problem. There are different methods to provide the cut-off point depending on the purpose of the classification, but there is a consensus that, without a clear reason to provide higher values for sensitivity or specificity, the Youden index provides an optimized balance of sensitivity/specificity [43].

Thus, the Youden index has been the most usual method to classify patients according to predictive or prognostic models in medicine. As previous studies provide methods to estimate the parameter of linear models in order to optimize ROC parameters, in this

work, we have proposed a stepwise algorithm that maximizes the Youden index. This algorithm is based on sequential optimizations, as they happen in dynamic programming, thus following the Bellman's optimality principle [44]. Unlike similar algorithms that use partial optimizations, we explore, at any step, the candidate biomarkers to be added to the model that provide linear combinations with the highest Youden index, following Pepe and Thompson's parameter search approach. In addition, our proposal also considers the ties that appear in the sequencing of the partial optimizations.

Our proposed approach has been explored in extensive simulation scenarios and compared with other methods. In particular, five other linear combination methods from the literature adapted to optimize the Youden index have been considered for comparison. Two of them are also based on Pepe and Thompson's empirical search (the Yin and Tian stepwise approach and the min-max approach) and three methods in numerical search based on derivatives (the classical logistic regression approach, a parametric approach and a non-parametric kernel smoothing approach).

The results obtained show that our proposed stepwise approach is superior to all other compared methods in most of the simulated scenarios considered for training data, but remains close to the rest for validation data, except in cases that are far from the verification of the normality hypothesis, in which, it is the best method for both training and validation data. It is globally followed by the Yin and Tian stepwise approach and the non-parametric kernel smoothing approach, the latter being slightly better in scenarios of higher correlation normal distributions for training data. However, in normal distributions scenarios, the non-parametric kernel smoothing approach outperformed Yian and Tian's stepwise approach for the validation data. In normal distribution scenarios, logistic regression and the parametric approach under multivariate normality showed a comparable performance overall, inferior to the non-parametric kernel smoothing approach in training data but superior or similar to the rest in validation data. However, the performance of the parametric approach worsened compared to logistic regression in non-normal distribution scenarios, as expected.

The min-max approach performed the worst in scenarios with different biomarker predictive capacities. However, it performed better in scenarios with the same predictive capacity of biomarkers (both in normal and non-normal distributions) and it outperformed the other algorithms when the covariance matrices differed between the diseased and non-diseased population. Among the wide range of simulated scenarios, highly negatively correlated biomarker scenarios were also included. In these scenarios, most algorithms achieved a very high performance, a result that is in agreement with the study by Pinsky and Zhu, who reported an increase in performance when considering highly negatively correlated biomarkers. In cases where they achieved near-perfect Youden indexes, the stepwise approaches performed worse than the other algorithms, with the exception of the min-max approach.

The performance of the linear combination methods was also analyzed on real datasets. The results obtained derived similar conclusions to those deduced from the simulated data. Remarkably, the stepwise approach performance is superior to the rest of the algorithms for the prostate cancer database in training and validation data. This is a data set where the PSA and free PSA variables for screening populations, or without previous treatment, present clear asymmetries that reflect the progression of the disease. This situation will occur for many other diseases where markers do not present results under the hypothesis of normality and the triggered values are associated with advanced stages of a disease. In these scenarios, the stepwise algorithm that we have proposed performs better than parametric algorithms where the non-verification of the hypothesis (normality or logistic relation) results in a loss of prediction capacity in the models.

The stepwise approaches and the non-parametric kernel smoothing approach achieved a good performance in general. Logistic regression also achieved one of the best discriminative capabilities on the DMD dataset (the best in the validation data), whose performance was relatively high overall.

Therefore, given the results of the spectrum analyzed, we could suggest the reader to use the min-max algorithm in scenarios with biomarkers with a similar predictive

power and different covariance matrices between the disease and non-disease group, and our proposed stepwise approach in other scenarios, especially for those apart from the normality hypothesis. In addition, we have created a library in R (***SLModels***) that can be used to implement these algorithms.

However, in terms of the computational time, stepwise approaches and, in particular, our proposal entail a significantly higher computational time due to the handling of ties, which may be more present in small sample sizes. This may be a limitation in the use of our algorithm, since, in practice, a faster computational speed is desired, especially when the number of biomarkers increases ($p > 5$). In these cases, the other algorithms (non-stepwise or min-max approach) have the advantage of being much more efficient. However, it has been shown that the min-max approach may not be sufficient in terms of discrimination in some scenarios. Aznar-Gimeno et al. [45] proposed a new approach that extends the min-max approach in order to analyze whether it increased predictive capacity while also being computationally tractable independently of the number of biomarkers.

As a line of future work, it is intended to optimize the proposed stepwise algorithm, with the aim of reducing its computational burden. It is intended to create tie handling strategies in such a way that the least pernicious criteria are used to break ties and not to drag them through many stages. The idea is to balance a certain increase in performance against an increase in computational load. Readers are also encouraged to adapt our algorithm using other target metrics to optimize and validate it in other scenarios, such as to explore and analyze the algorithm in multi-class classification problems using ROC surfaces.

5. Conclusions

In this work, we present a stepwise algorithm that complements and extends related existing ideas to optimize ROC-curve-derived parameters for linear models. We used, as the optimization parameter, the Youden index, which is the most used threshold point to dichotomize markers. As a strength, the developed method is a fully non-parametric distribution-free approach that showed a better performance in some scenarios. In addition, it captures the full predictive ability of a set of variables, in contrast to methodologies that try to reduce them. Additionally, the research has led to the creation of the R library ***SLModels***, which incorporates our proposed algorithm, can be used and is openly available to the scientific community. We believe that the findings of this research will provide insight for the development and application of algorithms for classification problems in medicine.

Author Contributions: Conceptualization, R.A.-G. and L.M.E.; methodology, R.A.-G., L.M.E. and G.S.; software, R.A.-G., L.M.E. and R.d.-H.-A.; validation, R.A.-G. and L.M.E.; formal analysis, R.A.-G. and L.M.E.; investigation, R.A.-G., L.M.E. and G.S.; resources, R.A.-G., L.M.E. and Á.B.-F.; data curation, R.A.-G., L.M.E. and Á.B.-F.; writing—original draft preparation, R.A.-G. and L.M.E.; writing—review and editing, R.A.-G., L.M.E., G.S., R.d.-H.-A. and Á.B.-F.; visualization, R.A.-G. and L.M.E.; supervision, R.A.-G., L.M.E., G.S. and R.d.-H.-A.; funding acquisition, G.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Government of Aragon (Stochastic Models Research group, grant number E46_20R), and Ministerio de Ciencia e Innovación (grant number PID2020-116873GB-I00).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Patient consent was waived due to the retrospective observational nature of this study; data could be fully anonymized.

Data Availability Statement: Data from the prostate cancer dataset were retrieved from the Miguel Servet University Hospital database and are not shared. The Duchenne muscular dystrophy dataset can be found at https://hbiostat.org/data/, accessed on 19 February 2022.

Acknowledgments: Research of Luis Mariano Esteban, Gerardo Sanz and Ángel Borque was supported by the Stochastic Models Research group of the Government of Aragon, grant number E46_20R and Ministerio de Ciencia e Innovación, grant number PID2020-116873GB-I00. In addition, we would like to thank Instituto Tecnológico de Aragón (ITAINNOVA), Zaragoza, Spain for the support of

the work of Rocío Aznar-Gimeno and Rafael del-Hoyo-Alonso, belonging to the Integration and Development of Big Data and Electrical Systems (IODIDE) group (T17_20R).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Esteban, L.M.; Sanz, G.; Borque, A. Linear combination of biomarkers to improve diagnostic accuracy in prostate cancer. *Monografías Matemáticas García de Galdeano* **2013**, *38*, 75–84.
2. Bansal, A.; Pepe, M.S. When does combining markers improve classification performance and what are implications for practice? *Stat. Med.* **2013**, *32*, 1877–1892. [CrossRef] [PubMed]
3. Yan, L.; Tian, L.; Liu, S. Combining large number of weak biomarkers based on AUC. *Stat. Med.* **2015**, *34*, 3811–3830. [CrossRef]
4. Lyu, T.; Ying, Z.; Zhang, H. A new semiparametric transformation approach to disease diagnosis with multiple biomarkers. *Stat. Med.* **2019**, *38*, 1386–1398. [CrossRef]
5. Amini, M.; Kazemnejad, A.; Zayeri, F.; Amirian, A.; Kariman, N. Application of adjusted-receiver operating characteristic curve analysis in combination of biomarkers for early detection of gestational diabetes mellitus. *Koomesh* **2019**, *21*, 751–758.
6. Ma, H.; Yang, J.; Xu, S.; Liu, C.; Zhang, Q. Combination of multiple functional markers to improve diagnostic accuracy. *J. Appl. Stat.* **2022**, *49*, 44–63. [CrossRef]
7. Yu, S. A Covariate-Adjusted Classification Model for Multiple Biomarkers in Disease Screening and Diagnosis. Ph.D. Thesis, Kansas State University, Manhattan, AR, USA, 2019.
8. Ahmadian, R.; Ercan, I.; Sigirli, D.; Yildiz, A. Combining binary and continuous biomarkers by maximizing the area under the receiver operating characteristic curve. *Commun. Stat. Simul. Comput.* **2020**, 1–14. [CrossRef]
9. Hu, X.; Li, C.; Chen, J.; Qin, G. Confidence intervals for the Youden index and its optimal cut-off point in the presence of covariates. *J. Biopharm. Stat.* **2021**, *31*, 251–272. [CrossRef]
10. Kang, L.; Xiong, C.; Crane, P.; Tian, L. Linear combinations of biomarkers to improve diagnostic accuracy with three ordinal diagnostic categories. *Stat. Med.* **2013**, *32*, 631–643. [CrossRef]
11. Maiti, R.; Li, J.; Das, P.; Feng, L.; Hausenloy, D.; Chakraborty, B. A distribution-free smoothed combination method of biomarkers to improve diagnostic accuracy in multi-category classification. *arXiv* **2019**, arXiv:1904.10046.
12. Su, J.Q.; Liu, J.S. Linear combinations of multiple diagnostic markers. *J. Am. Stat. Assoc.* **1993**, *88*, 1350–1355. [CrossRef]
13. Pepe, M.S.; Thompson, M.L. Combining diagnostic test results to increase accuracy. *Biostatistics* **2000**, *1*, 123–140. [CrossRef] [PubMed]
14. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36. [CrossRef] [PubMed]
15. Liu, C.; Liu, A.; Halabi, S. A min–max combination of biomarkers to improve diagnostic accuracy. *Stat. Med.* **2011**, *30*, 2005–2014. [CrossRef] [PubMed]
16. Pepe, M.S.; Cai, T.; Longton, G. Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics* **2006**, *62*, 221–229. [CrossRef]
17. Esteban, L.M.; Sanz, G.; Borque, A. A step-by-step algorithm for combining diagnostic tests. *J. Appl. Stat.* **2011**, *38*, 899–911. [CrossRef]
18. Kang, L.; Liu, A.; Tian, L. Linear combination methods to improve diagnostic/prognostic accuracy on future observations. *Stat. Methods Med. Res.* **2016**, *25*, 1359–1380. [CrossRef]
19. Liu, A.; Schisterman, E.F.; Zhu, Y. On linear combinations of biomarkers to improve diagnostic accuracy. *Stat. Med.* **2005**, *24*, 37–47. [CrossRef]
20. Yin, J.; Tian, L. Joint inference about sensitivity and specificity at the optimal cut-off point associated with Youden index. *Comput. Stat. Data Anal.* **2014**, *77*, 1–13 [CrossRef]
21. Yu, W.; Park, T. Two simple algorithms on linear combination of multiple biomarkers to maximize partial area under the ROC curve. *Comput. Stat. Data Anal.* **2015**, *88*, 15–27 [CrossRef]
22. Yan, Q.; Bantis, L.E.; Stanford, J.L.; Feng, Z. Combining multiple biomarkers linearly to maximize the partial area under the ROC curve. *Stat. Med.* **2018**, *37*, 627–642. [CrossRef] [PubMed]
23. Ma, H.; Halabi, S.; Liu, A. On the use of min-max combination of biomarkers to maximize the partial area under the ROC curve. *J. Probab. Stat.* **2019**, *2019*, 8953530. [CrossRef] [PubMed]
24. Perkins, N.J.; Schisterman, E.F. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *Am. J. Epidemiol.* **2006**, *163*, 670–675. [CrossRef] [PubMed]
25. Youden, W.J. Index for rating diagnostic tests. *Cancer J.* **1950**, *3*, 32–35. [CrossRef]
26. Martínez-Camblor, P.; Pardo-Fernández, J.C. The Youden Index in the Generalized Receiver Operating Characteristic Curve Context. *Int. J. Biostat.* **2019**, *15*, 20180060. [CrossRef]
27. Yin, J.; Tian, L. Optimal linear combinations of multiple diagnostic biomarkers based on Youden index. *Stat. Med.* **2014**, *33*, 1426–1440. [CrossRef]
28. Yin, J.; Tian, L. Joint confidence region estimation for area under ROC curve and Youden index. *Stat. Med.* **2014**, *33*, 985–1000. [CrossRef]

29. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020. Available online: http://www.r-project.org/index.html (accessed on 19 February 2022).
30. SLModels: Stepwise Linear Models for Binary Classification Problems under Youden Index Optimisation. R Package Version 0.1.2. Available online: https://cran.r-project.org/web/packages/SLModels/index.html (accessed on 19 February 2022).
31. Walker, S.H.; Duncan, D.B. Estimation of the probability of an event as a function of several independent variables. *Biometrika* **1967**, *54*, 167–179. [CrossRef]
32. Schisterman, E.F.; Perkins, N. Confidence intervals for the Youden index and corresponding optimal cut-point. *Commun. Stat. Simul. Comput.* **2007**, *36*, 549–563. [CrossRef]
33. Faraggi, D.; Reiser, B. Estimation of the area under the ROC curve. *Stat. Med.* **2002**, *21*, 3093–3106. [CrossRef]
34. Rosenblatt, M. Remarks on some nonparametric estimates of a density function. *Ann. Math. Stat.* **1956**, *27*, 832–837. [CrossRef]
35. Parzen, E. On estimation of a probability density function and mode. *Ann. Math. Stat.* **1962**, *33*, 1065–1076. [CrossRef]
36. Fluss, R.; Faraggi, D.; Reiser, B. Estimation of the Youden Index and its associated cutoff point. *Biom. J. J. Math. Biol.* **2005**, *47*, 458–472. [CrossRef] [PubMed]
37. Silverman, B.W. *Density Estimation for Statistics and Data Analysis*; Routledge: London, UK, 2018.
38. Percy, M.E.; Andrews, D.F.; Thompson, M.W. Duchenne muscular dystrophy carrier detection using logistic discrimination: Serum creatine kinase, hemopexin, pyruvate kinase, and lactate dehydrogenase in combination. *Am. J. Med. Genet. A* **1982**, *13*, 27–38. [CrossRef] [PubMed]
39. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef]
40. Rubio-Briones, J.; Borque-Fernando, A.; Esteban, L.M.; Mascarós, J.M.; Ramírez-Backhaus, M.; Casanova, J.; Collado, A.; Mir, C.; Gómez-Ferrer, A.; Wong, A.; et al. Validation of a 2-gene mRNA urine test for the detection of >=GG2 prostate cancer in an opportunistic screening population. *Prostate* **2020**, *80*, 500–507. [CrossRef]
41. Morote, J.; Schwartzman, I.; Borque, A.; Esteban, L.M.; Celma, A.; Roche, S.; de Torres, I.M.; Mast, R.; Semidey, M.E.; Regis, L.; et al. Prediction of clinically significant prostate cancer after negative prostate biopsy: The current value of microscopic findings. In *Urologic Oncology: Seminars and Original Investigations*; Elsevier: Amsterdam, The Netherlands, 2020.
42. Pinsky, P.F.; Zhu, C.S. Building Multi-Marker Algorithms for Disease Prediction—The Role of Correlations among Markers. *Biomark. Insights* **2011**, *6*, 83–93. [CrossRef]
43. Rota, M.; Antolini, L. Finding the optimal cut-point for Gaussian and Gamma distributed biomarkers. *Comput. Stat. Data Anal.* **2014**, *69*, 1–14. [CrossRef]
44. Bellman, R.E. *Dynamic Programming*; Princeton University Press: Princeton, NJ, USA, 1957.
45. Aznar-Gimeno, R.; Esteban, L.M.; Sanz, G.; del-Hoyo-Alonso, R.; Savirón-Cornudella, R.; Antolini, L. Incorporating a New Summary Statistic into the Min–Max Approach: A Min–Max–Median, Min–Max–IQR Combination of Biomarkers for Maximising the Youden Index. *Mathematics* **2021**, *9*, 2497. [CrossRef]

Capítulo 6

Enfoques Min-Max-Median, Min-Max-IQR bajo maximización del índice de Youden

En este capítulo se presentan los trabajos [8, 7], siguiendo la línea de investigación de estimación de modelos de clasificación con criterios de optimalidad derivados de la curva ROC. En concreto, en el trabajo [8] se proponen nuevos enfoques no paramétricos (enfoque Min-Max-Median, Min-Max-IQR) para la combinación de biomarcadores bajo la maximización del índice de Youden. En el trabajo [7] se presenta una comparación de los enfoques propuestos con otros algoritmos del estado del arte, que incluyen el algoritmo de Machine Learning (ML) XGBoost. En las siguientes secciones se explica el propósito de estas investigaciones y las aportaciones de cada trabajo, [8] y [7] respectivamente.

6.1. Enfoques propuestos: Min-Max-Median/IQR

Siguiendo la línea de estudio del capítulo previo, Liu et al. [56] sugieren un enfoque no paramétrico, denominado enfoque min-max, que también aborda la limitación computacional de los estudios de la literatura. Este enfoque tiene una carga computacional más eficiente que los enfoques paso a paso, puesto que se basa en combinar linealmente los valores mínimo y máximo de los biomarcadores, involucrando una sola búsqueda de coeficientes (sección 4.1.3). Esto supone una ventaja ya que, independientemente del número de biomarcadores originales, su carga computacional es similar. Sin embargo, se ha demostrado en diversos estudios, particularmente en nuestro trabajo [4] presentado en el capítulo anterior, que existen escenarios en los que el enfoque min-max alcanza una optimalidad notablemente inferior a otros métodos como los algoritmos paso a paso. Esto indica que la combinación de los valores mínimo y máximo (estadísticos extremos) podría no ser suficiente en términos de discriminación.

La motivación de este trabajo [8] se basa en estos resultados. Proponemos nuevos enfoques, denominados enfoque Min-Max-Median (MMM) y enfoque Min-Max-IQR (MMIQR), que amplían el enfoque propuesto por Liu et al. [56], incorporando una nueva estadística de resumen, que incluye el valor central (mediana) o una medida de dispersión (rango intercuartílico). Esto permite considerar una mayor información de los biomarcadores originales. Nuestro objetivo fue mejorar la capacidad predictiva, al mismo tiempo que se mantenía una carga computacional abordable, independientemente del número de biomarcadores. En concreto, nuestro enfoque implica la estimación de la combinación lineal óptima de tres variables. Para seleccionar esta combinación óptima, utilizamos nuestro enfoque paso a paso propuesto en el trabajo [4], presentado en el capítulo anterior. La elección de este enfoque se fundamenta en su rendimiento, que ha demostrado tener una de las mejores capacidades discriminatorias, con índices de Youden más altos, tanto en los datos reales como en simulaciones fuera de la normalidad, que suelen ser escenarios comunes en la práctica. Esto lo convierte en una opción valiosa para nuestro propósito.

Los enfoques propuestos (MMM, MMIQR) fueron analizados y comparados con el método min-max y la regresión logística, en diferentes escenarios de datos simulados y datos reales. El objetivo era comparar su rendimiento con el enfoque que ampliamos (método min-max) y el método tradicional de regresión logística, ambos eficientes computacionalmente. Los datos simulados analizados cubren una amplia variedad de escenarios en términos de la distribución de biomarcadores, la capacidad para discriminar de los biomarcadores y la correlación entre ellos, teniendo en cuenta tamaños de muestra desde pequeños hasta más grandes. En cuanto a los conjuntos de datos reales, se consideraron datos para el diagnóstico de la distrofia muscular de Duchenne y la predicción de SGA (Small for Gestational Age, en inglés).

Los resultados muestran que nuestras propuestas superan a la regresión logística clásica en escenarios de biomarcadores con la misma capacidad predictiva y diferentes matrices de covarianza entre grupos. Nuestros enfoques funcionan mejor que el enfoque min-max cuando los biomarcadores son independientes y en escenarios reales.

En conclusión, en este estudio presentamos nuevos enfoques no paramétricos para la combinación de biomarcadores que buscan optimizar el índice de Youden. Estos enfoques aprovechan las ventajas de los estudios previos en la literatura, al mismo tiempo que abordan sus limitaciones, con el objetivo de mejorar su capacidad predictiva.

La investigación realizada, junto con la presentada en el capítulo anterior [4], ha resultado también en la creación de la librería SLModels [6], que proporciona acceso a los algoritmos propuestos (enfoque paso a paso, enfoque MMM, enfoque MMIQR), así

como al enfoque min-max, para su libre utilización por parte de la comunidad científica. La información detallada sobre el uso de la librería SLModels se encuentra disponible en el Anexo A.

A continuación, se presenta el artículo que describe el trabajo realizado, donde se detallan los enfoques propuestos (MMM, MMIQR) y se discute en mayor profundidad la investigación llevada a cabo.

mathematics

MDPI

Article

Incorporating a New Summary Statistic into the Min–Max Approach: A Min–Max–Median, Min–Max–IQR Combination of Biomarkers for Maximising the Youden Index

Rocío Aznar-Gimeno [1,*], Luis M. Esteban [2,*], Gerardo Sanz [3], Rafael del-Hoyo-Alonso [1] and Ricardo Savirón-Cornudella [4]

[1] Department of Big Data and Cognitive Systems, Instituto Tecnológico de Aragón (ITAINNOVA), 50018 Zaragoza, Spain; rdelhoyo@itainnova.es
[2] Department of Applied Mathematics, Escuela Universitaria Politécnica de La Almunia, Universidad de Zaragoza, La Almunia de Doña Godina, 50100 Zaragoza, Spain
[3] Department of Statistical Methods and Institute for Biocomputation and Physics of Complex Systems-BIFI, University of Zaragoza, 50009 Zaragoza, Spain; gerardo.sanz@unizar.es
[4] Department of Obstetrics and Gynecology, Hospital Clínico San Carlos and Instituto de Investigación Sanitaria San Carlos (IdISSC), Universidad Complutense de Madrid, 28040 Madrid, Spain; rsaviron@gmail.com
* Correspondence: raznar@itainnova.es (R.A.-G.); lmeste@unizar.es (L.M.E.)

Citation: Aznar-Gimeno, R.; Esteban, L.M.; Sanz, G.; del-Hoyo-Alonso, R.; Savirón-Cornudella, R. Incorporating a New Summary Statistic into the Min–Max Approach: A Min–Max–Median, Min–Max–IQR Combination of Biomarkers for Maximising the Youden Index. *Mathematics* **2021**, *9*, 2497. https://doi.org/10.3390/math9192497

Academic Editors: Sorana D. Bolboacă and Polychronis Kostoulas

Received: 30 August 2021
Accepted: 30 September 2021
Published: 5 October 2021

Abstract: Linearly combining multiple biomarkers is a common practice that can provide a better diagnostic performance. When the number of biomarkers is sufficiently high, a computational burden problem arises. Liu et al. proposed a distribution-free approach (min–max approach) that linearly combines the minimum and maximum values of the biomarkers, involving only a single coefficient search. However, the combination of minimum and maximum biomarkers alone may not be sufficient in terms of discrimination. In this paper, we propose a new approach that extends that of Liu et al. by incorporating a new summary statistic, specifically, the median or interquartile range (min–max–median and min–max–IQR approaches) in order to find the optimal combination that maximises the Youden index. Although this approach is more computationally intensive than the one proposed by Liu et al, it includes more information and the number of parameters to be estimated remains reasonable. We compare the performance of the proposed approaches (min–max–median and min–max–IQR) with the min–max approach and logistic regression. For this purpose, a wide range of different simulated data scenarios were explored. We also apply the approaches to two real datasets (Duchenne Muscular Dystrophy and Small for Gestational Age).

Keywords: linear combination; biomarkers; classification; Youden index; summary statistic; min–max approach; logistic regression; median; interquartile range

1. Introduction

In clinical practice, it is common to have information on multiple biomarkers for disease diagnosis. Combining them all into a single marker is common practice and usually offers better diagnostic accuracy than considering each biomarker separately [1,2]. How to increment the performance of a standard marker by adding new markers has also received attention in the literature [3].

The combination of continuous biomarkers provides a continuous value and the dichotomisation of this value is important as it provides the clinician with a classification rule related to a disease and their clinical decisions are usually based on such categorisation of patients into groups [4].

Although different algorithms have been considered to provide predictive models, linear models (linear combination of biomarkers) have been widely proposed for binary classification problems due to their simplicity of interpretation and good performance [5].

The accuracy of the diagnostic marker is usually assessed on the basis of its discriminatory ability as analysed by the Receiver Operating Characteristic (ROC) curve through derived statistics such as sensitivity and specificity pairs, the area or partial area under the ROC curve (AUC/pAUC), or the Youden index [6]. Specifically, the formulation of algorithms for the estimation of binary classification models that maximise the AUC has been widely explored and has been the background and basis for many subsequent algorithms recently published [7–9].

Su and Liu [10] provided the estimation of linear models that maximises AUC under multivariate normality. However, this normality assumption is a high-demand hypothesis on real data and is often not easy to meet in clinical reality. This real limitation leads to the need for more flexible statistical approaches that are not subject to distributional assumptions [11,12]. Several statistical methods have been presented in the literature that do not depend on a specific distribution. One example is the statistics of the fractional moments, whose effectiveness and advantages have been demonstrated in real scenarios [13,14]. Pepe et al. [15,16] proposed a distribution-free approach to estimate the linear model that maximise AUC based on the Mann–Whitney U statistic [17], in order to address the Su and Liu limitation. The process that Pepe et al. proposed is a discrete optimisation which is based on an extensive search over the parameter vector. However, this process is not computationally accessible (NP-hard problem) for a number of biomarkers $p \geq 3$. In order to make it computationally tractable, Pepe et al. also suggested the use of stepwise algorithms that are an alternative which have shown to perform well in various scenarios. The idea of these algorithms is to include a new biomarker at each step by selecting the best combination of two biomarkers. The approach and suggestions proposed by Pepe et al. have been the source of development of non-parametric and semiparametric approaches in the construction of classifiers under optimality criteria derived from the ROC curve.

Esteban et al. [18] implemented the approach of Pepe et al. providing strategies to handle ties that appear in the sequencing of partial optimisations under AUC optimisation criteria. Kang et al. [19,20] proposed a less computationally demanding stepwise combination approach based on a rated of AUC values that correspond to predictors variables. However, finding the optimal parameters in these stepwise approaches becomes a computational burden problem that is also difficult to tackle when the number of biomarkers is high ($p > 5$).

Liu et al. [21] proposed an approach, called the min–max approach, which linearly combines the minimum and maximum values of biomarkers, involving a single coefficient search that maximises the Mann–Whitney U statistic of AUC. The advantage of this approach is that it is computationally tractable regardless of the number of biomarkers. However, it has been shown that there are scenarios where the min–max approach generally achieves less optimality compared to other methods that use information from all biomarkers, such as stepwise approaches [19,22]. This may indicate that the combination of minimum and maximum biomarkers alone may not be sufficient in terms of discrimination [5].

Algorithms focused on optimising other parameters derived from the ROC curve have also been developed. Liu et al. [23] analysed the optimal linear combination into diagnostic marker maximising sensitivity over a range of specificity. Yin and Tian [24] proposed approaches to estimate the joint confidence region of sensitivity and specificity at the cut-off point determined by the Youden index. Based on the stepwise combination approach of Kang et al., Yin and Tian [22] conducted a study with the aim of optimising the Youden index. Yin and Tian [25] also studied the optimisation of the AUC and the Youden index simultaneously and presented both parametric and non-parametric approaches that estimate the joint confidence region of the AUC and the Youden index. Based on the min–max approach, Ma et al. [26] adapted and extended the min–max method to the estimation of the pAUC for multiple continuous biomarkers. Yu and Park [27] and Yan et al. [28] also explored methods for the linear combination of multiple biomarkers that optimise the pAUC.

The Youden index is a statistical metric widely and successfully used in several clinical studies and, at the same time, serves as an appropriate summary for making the diagnosis and as a criterion for choosing the best cut-off points for dichotomising a biomarker [29], which is key for the diagnosis of the disease. There are different methods for providing the best cut-off point depending on the aim of the classification, but there is consensus that, without a clear reason to provide higher values for sensitivity or specificity, the Youden index provides an optimised balance of sensitivity/specificity.

In order to address the computational burden regardless of the number of biomarkers and to increase the predictive capacity, in this work we propose a new approach that extends that of Liu et al. by incorporating a new summary statistic, in particular, the median or interquartile range (min–max–median and min–max–IQR approaches). Although this approach is more computationally intensive than the one proposed by Liu et al., it includes more information and the number of parameters to estimate remains reasonable. For the choice of the best linear combination of these three variables, we use a stepwise method (two steps) following the empirical search approach of Pepe and Thompson. For the search of the optimal linear combination, we consider the Youden index as an objective function.

We compared the performance of the proposed algorithm (min–max–median and min–max–IQR approaches) with the min–max approach adapted to optimise the Youden index and logistic regression, which are computationally tractable linear approaches. For this purpose, a comprehensive study was carried out considering different simulated data scenarios. The performances of the approaches on real datasets (e.g., Duchenne muscular dystrophy and Small for Gestational Age prediction) were also analysed. The comparison of the performance of the algorithms was carried out by considering the results obtained on the entire dataset (resubstitution method) and using a validation procedure derived from K-fold cross-validation, which evaluates the generalisability of the predictive model and avoids possible overfitting [30].

Thus, the aim of this work is to propose a new approach (min–max–median/min–max–IQR approach) that has the advantage of being computable regardless of the number of biomarkers and that does not depend on any distributional assumptions, and analyse their performance in comparison with the min–max approach and the logistic regression which is a standard in prediction models with good performance, to find optimal scenarios.

2. Materials and Methods

Firstly, we introduce the formulation of the linear model and estimation suggestions of Pepe et al. [15,16]. This is the basis for the formulation and estimation of the linear model of summary statistic-based approaches. Then, the methods based on Youden index maximisation to be compared are presented: the min–max approach, the min–max–median and min–max–IQR approaches and logistic regression. Finally, the simulated scenarios as well as the real datasets considered are described. All methods were programmed and applied using the free software R [31].

2.1. Background: Pepe et al.'s Approach

Pepe and Thompson [15] proposed a non-parametric approach in order to estimate the linear model that maximises the AUC based on the Mann–Whitney U statistic [17]. Thus, this approach allows its use for all types of data without any distribution assumptions. The formulation of the linear model is as follows:

$$L(\mathbf{X}) = X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \quad (1)$$

where p denotes the number of biomarkers, X_i the biomarker $i \in [1, \ldots, p]$ and β_i the parameter to be estimated. If, in addition, we assumed that we have n_k individuals of the group $k = 1, 2$ (diseased and non-diseased, respectively) and $\mathbf{X_{ki}}$ is the vector of

p biomarkers for the i^{th} individual of group k, then the empirical AUC based on the Mann–Whitney U statistic is given by the following expression:

$$\widehat{AUC} = \frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} I(L(\mathbf{X_{1i}}) > L(\mathbf{X_{2j}})) + \frac{1}{2} I(L(\mathbf{X_{1i}}) = L(\mathbf{X_{2j}}))}{n_1 \cdot n_2} \tag{2}$$

For the estimation of the parameter β_i, Pepe et al. suggested a discrete optimisation which is based on a grid search over 201 equally spaced values in the interval $[-1, 1]$. For simplicity, consider the linear combination of two biomarkers: $X_i + \beta X_j$. Due to the invariant property of the ROC curve for any monotonic transformation, dividing by the β value does not change the value of the sensitivity and specificity pair. Thus, estimating $X_i + \beta X_j$ for $\beta > 1$ and $\beta < -1$ is equivalent to estimating $\frac{1}{\beta} X_i + X_j$ for $\beta \in [-1, 1]$ and, therefore, all possible values of $\beta \in \mathbb{R}$ are covered.

However, this optimisation requires a large computational effort for dimensions $p \geq 3$. In order to make it computationally accessible, Pepe et al. suggested using stepwise algorithms in which a new biomarker is included at each stage, selecting the best combination of two biomarkers. This turns an unaddressable computational burden problem into an addressable one by applying $p - 1$ times Pepe et al.'s suggested single-parameter estimate.

Although Pepe et al. originally introduced this approach to optimise the AUC, both the formulation of the linear model (1), the suggested empirical search and the suggested use of stepwise approaches are the basis in the development of the min–max approaches and our proposals: min–max–median and min–max–IQR approaches under Youden index maximisation.

Let X_{kij} be the j^{th} biomarker ($j = 1, \ldots, p$) for the i^{th} individual of group k = 1,2 (disease and non-disease, respectively), $\mathbf{X_{ki}}$ the vector of p biomarkers for the i^{th} individual of group k and $\mathbf{X_k} = \left(\mathbf{X_{k1}}, \ldots, \mathbf{X_{kn_k}}\right)$ where n_k is the number of individuals in group k. The Youden Index (J) of the linear combination is defined as:

$$\begin{aligned} J &= \max_c \{Sensitivity(c) + Specificity(c) - 1\} \\ &= \max_c \{F_{\mathbf{Y_2}}(c) - F_{\mathbf{Y_1}}(c)\} \end{aligned} \tag{3}$$

where c denotes the cut-off point and $F_{\mathbf{Y_k}}(c) = P(\mathbf{Y_k} \leq c)$ the cumulative distribution function of random variable $\mathbf{Y_k} = \boldsymbol{\beta}^T \mathbf{X_k}, k = 1, 2$ (linear combination), $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ being the parameter vector to be estimated. It is therefore immediate to deduce the following formula from the empirical estimation of the Youden index:

$$\begin{aligned} \hat{J}_{\mathbf{fi}} &= \hat{F}_{\mathbf{Y_2}}(\hat{c}_{\boldsymbol{\beta}}) - \hat{F}_{\mathbf{Y_1}}(\hat{c}_{\boldsymbol{\beta}}) \\ &= \frac{\sum_{i=1}^{n_2} I\left(\boldsymbol{\beta}^T \mathbf{X_{2i}} \leq \hat{c}_{\boldsymbol{\beta}}\right)}{n_2} - \frac{\sum_{i=1}^{n_1} I\left(\boldsymbol{\beta}^T \mathbf{X_{1i}} \leq \hat{c}_{\boldsymbol{\beta}}\right)}{n_1} \end{aligned} \tag{4}$$

where $c_{\boldsymbol{\beta}} = \{c : \max_c \left(F_{\mathbf{Y_2}}(c) - F_{\mathbf{Y_1}}(c)\right)\}$ denotes the optimal cut-off point.

2.2. Min–Max Approach

Liu et al. [21] proposed the distribution-free min–max approach. The idea of this approach is based on reducing the order of the linear combination by considering only two markers from the measurements information of the original p biomarkers (maximum value and the minimum value of all the p biomarkers):

$$X_{\max} + \beta X_{\min} \tag{5}$$

Therefore, this approach is reduced to estimating a single β parameter of the linear combination that maximises the AUC, as it is based on the linear model proposed by Pepe et al. expressed in Equation (1).

This approach was adapted in our work with in order to maximise the Youden index with the following expression:

$$\hat{J}_\beta = \frac{\sum_{i=1}^{n_2} I\left(X_{2i,\max} + \beta X_{2i,\min} \leq \hat{c}_\beta\right)}{n_2} - \frac{\sum_{i=1}^{n_1} I\left(X_{1i,\max} + \beta X_{1i,\min} \leq \hat{c}_\beta\right)}{n_1} \quad (6)$$

where $X_{ki,\max} = \max_{1\leq j\leq p}\left(X_{kij}\right)$ and $X_{ki,\min} = \min_{1\leq j\leq p}\left(X_{kij}\right)$ for each $i = 1, \ldots, n_k$. $k = 1, 2$.

The search for the optimal parameter follows the empirical search suggested by Pepe et al.: for each value β of the 201 equally spaced $\in [-1, 1]$, the optimal cut-off point $(\hat{c}_\beta)$ that maximises Youden index is calculated. The final value chosen $(\hat{\beta})$ is the one with the highest Youden $(\hat{J}_\beta)$ obtained.

2.3. Our Proposed Summary Statistics-Based Approaches: Min–Max–Median and Min–Max–IQR

We propose two new approaches based on summary statistics to maximise the Youden index. The underlying idea of the approaches is based on the same idea as the min–max approach (Section 2.2) of reducing the dimension of the problem by considering summary statistic information of the p biomarkers as variables. They could be seen as an extension of the min–max approach by incorporating a new summary statistic. The aim of this proposal was to incorporate more information from the original biomarkers while keeping the approach computationally tractable.

Specifically, we propose two approaches: including the median and IQR at the maximum and minimum of the p biomarkers measurements. Therefore, our approach is reduced to considering the linear combination of three variables (max, min, median–max, min, IQR).

For the estimation of the optimal linear combination of these three variables, a stepwise approach is used, as suggested by Pepe et al., in order to ensure an approachable computational time. Specifically, our proposals implement an adaptation of the stepwise approach proposed by Esteban et al. [18] for Youden Index maximisation, where at each step a new variable is introduced and the optimal linear combination of two variables is selected using the empirical search suggested by Pepe et al. In this way, we turn a time-consuming problem into a computationally tractable problem by considering two steps of optimising the linear combination of two variables.

The proposed approach is explained in detail in the following steps. Consider the min–max–median approach. The min–max–IQR approach is equivalent:

1. Firstly, the problem of estimating the optimal linear combination of p variables is transformed to the estimation of the optimal linear combination of three variables (min, max, median). For this purpose, the values of these new variables are calculated for each i individual from their values of the p original biomarkers:

$$X_{ki,\max} = \max_{1\leq j\leq p}\left(X_{kij}\right),\ X_{ki,\min} = \min_{1\leq j\leq p}\left(X_{kij}\right),\ X_{ki,median} = median_{1\leq j\leq p}\left(X_{kij}\right) \text{ for each } i = 1, \ldots, n_k.,\ k = 1, 2. \quad (7)$$

2. Once the information is transformed, a stepwise approach is used to estimate the optimal linear combination of the three new variables (min–max–median). For simplicity, denote by X_{kij} the values of the new transformed variables (min–max–median, $j = 1, 2, 3$) for the i^{th} individual of group k = 1, 2. The first step of the stepwise approach consists of choosing the combination of two variables that maximises the Youden index, using empirical search proposed by Pepe et al.:

$$\hat{J}_{\beta_2} = \frac{\sum_{i=1}^{n_2} I\left(X_{2ij} + \beta_2 X_{2il} \leq \hat{c}_{\beta_2}\right)}{n_2} - \frac{\sum_{i=1}^{n_1} I\left(X_{1ij} + \beta_2 X_{1il} \leq \hat{c}_{\beta_2}\right)}{n_1} \quad \beta_2 \in [-1, 1],\ \ \forall j \neq l = 1, 2, 3 \quad (8)$$

The maximum Youden index can be reached for more than one optimal linear combination. Our approach considers these ties and drags them forward in the next

step. For simplicity, consider that it is reached for a single linear combination. The optimal combination chosen (two variables and one parameter β_2) that maximises the Youden index (8) is considered as a single variable in the next step (and final) of the stepwise approach. For simplicity, suppose the following optimal linear combination: $X_{ki1} + \beta_2 X_{ki2}$.

3. Finally, the last variable (X_{ki3}) not chosen in the previous step is included and the optimal linear combination of the two variables is chosen as in point 2. Specifically, either combination (9) or (10) that maximises the Youden index $\hat{J}_{\beta_3}$ is selected:

$$\hat{J}_{\beta_3} = \frac{\sum_{i=1}^{n_2} I\left((X_{2i1} + \beta_2 X_{2i2}) + \beta_3 X_{2i3} \leq \hat{c}_{\beta_3}\right)}{n_2} - \frac{\sum_{i=1}^{n_1} I\left((X_{1i1} + \beta_2 X_{1i2}) + \beta_3 X_{1i3} \leq \hat{c}_{\beta_3}\right)}{n_1} \quad \beta_3 \in [-1, 1] \tag{9}$$

$$\hat{J}_{\beta_3} = \frac{\sum_{i=1}^{n_2} I\left(\beta_3(X_{2i1} + \beta_2 X_{2i2}) + X_{2i3} \leq \hat{c}_{\beta_3}\right)}{n_2} - \frac{\sum_{i=1}^{n_1} I\left(\beta_3(X_{1i1} + \beta_2 X_{1i2}) + X_{1i3} \leq \hat{c}_{\beta_3}\right)}{n_1} \quad \beta_3 \in [-1, 1] \tag{10}$$

The code of all approaches is available upon request to Ms. Rocío Aznar-Gimeno (raznar@itainnova.es).

The predictive ability of these approaches based on summary statistics (min–max approach and min–max–median/IQR approach) is also compared with the classical logistic regression approach [32], which is a subtype of generalised linear model to predict the probability of an event (disease or non-disease) given a set of variables through the logistic function. We used for comparisons the logistic regression method in the multivariate binormal setting, for the diseased and non-diseased populations, which showed similar performances in the linear discriminant analysis, which is standard in predictive models, exhibiting a good performance in all scenarios [15]. The objective was to compare computationally tractable linear combination methods, regardless of the number of original biomarkers, dealing with biomarkers' real values.

2.4. Performance Comparison

The comparison was carried out by considering both the resubstitution method and a method from 5-fold cross-validation. For this purpose, a range of simulated data were explored. Their application to two real dataset examples was also explored in order to evaluate their performances in real cases.

2.4.1. Simulations

The simulated data analysed cover a wide range of scenarios in terms of the distribution of biomarkers, the ability to discriminate between biomarkers and the correlation between them, considering smaller to larger sample sizes.

In order to observe behaviours with smaller and larger numbers of biomarkers, four biomarkers ($p = 4$) and ten biomarkers ($p = 10$) were considered in the simulated scenarios. The peculiarity of statistics-based approaches (in particular our proposed approaches) is that they translate biomarker information into only 3 variables, regardless of the number of initial biomarkers. This has the advantage that they are always computationally feasible. Furthermore, in scenarios where the number of biomarkers to be considered exceeds the sample size, these approaches can avoid overfitting to the data, due to their inherent dimensionality reduction characteristics. In order to analyse the performance of the algorithms in these specific cases, simulated scenarios of $p = 100$ biomarkers and a smaller sample size ($n_1 = n_2 = 30$ were also explored.

Different joint distributions were considered following multivariate normal ($\mathbf{X_k} \sim MVN(\mathbf{m_k}, \mathbf{\Sigma_k})$) and also non-normal distributions (in particular, log-normal: $e^{\mathbf{X_k}}$) in both diseased and non-diseased populations ($k = 1, 2$). In this way, both normal distributions, which is one of the common distributions in statistics, and distributions that allow assessment beyond normality and symmetry, were analysed.

When the multivariate normal distribution for diseased and non-diseased population are assumed $\mathbf{X} \sim N(\mathbf{m_1}, \mathbf{\Sigma_1})$, $\mathbf{Y} \sim N(\mathbf{m_2}, \mathbf{\Sigma_2})$ the coefficients for the optimal linear combi-

nation (which coincides with the Linear Discrimination Function (LDF)) and its area under the ROC curve are known and given by $\beta_{max} \propto (\mathbf{\Sigma_1} + \mathbf{\Sigma_2})^{-1}\mathbf{m}$, where $\mathbf{m} = \mathbf{m_1} - \mathbf{m_2}$ and $\mathrm{AUC}_{\max} = \Phi(\sqrt{\mathbf{m}^{\mathrm{T}}(\mathbf{\Sigma_1} + \mathbf{\Sigma_2})^{-1}\mathbf{m}})$ (see Su and Liu [10]).

Using this formula, we simulated markers by choosing mean vectors and variance–covariance matrices, whose AUC shown a good discrimination ability, with an AUC above 0.8 or at least near 0.8. The reason behind this is that these models are those whose dichotomisation by Youden index is interesting in real practice.

Regarding the vector of means, the null vector was considered for the non-diseased population ($\mathbf{m_2} = \vec{0}$) in all simulated scenarios. For the diseased population, scenarios considering both the same and different abilities to discriminate between biomarkers were explored: (i) same mean for each biomarker: $\mathbf{m_1} = (1.0, 1.0, \ldots, 1.0)^T$; (ii) different means: $\mathbf{m_1} = (0.2, 0.5, 1.0, 0.7)^T$, $\mathbf{m_1} = (0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0)^T$.

Regarding the variance–covariance matrix, for simplicity, the variance of each biomarker was set to 1, so that covariances equalled correlations. Scenarios were analysed with the same covariance matrix between populations ($\mathbf{\Sigma_1} = \mathbf{\Sigma_2}$) and different ($\mathbf{\Sigma_1} \neq \mathbf{\Sigma_2}$). Specifically, for the cases with the same covariance matrix, the following scenarios were explored: (i) independence ($\mathbf{\Sigma_1} = \mathbf{\Sigma_2} = I$), (ii) medium correlation ($\mathbf{\Sigma_1} = \mathbf{\Sigma_2} = 0.5 \cdot I + 0.5 \cdot J$), where I is the identity matrix and J a matrix of all ones, and (iii) negative correlation ($\rho = -0.1$) for all pairs of biomarkers. As for the scenarios with different covariance matrices between populations, the following scenarios were analysed: (i) $\mathbf{\Sigma_1} = 0.3 \cdot I + 0.7 \cdot J$ (high correlation) and $\mathbf{\Sigma_2} = 0.7 \cdot I + 0.3 \cdot J$ (low correlation) (ii) $\mathbf{\Sigma_1} = 0.5 \cdot I + 0.5 \cdot J$ (medium correlation) and $\mathbf{\Sigma_2} = I$ (independents).

For each scenario, 1000 random samples from underlying distribution were considered with the following different samples sizes: (i) $n_1 = n_2 = 30$; (ii) $n_1 = n_2 = 100$; (iii) $n_1 = n_2 = 500$. Each method was applied to each of the simulated scenarios and the maximum Youden index was obtained.

2.4.2. Application to Real Data

The analysed methods were applied to two real data examples with the aim of diagnosing and predicting of Duchenne Muscular Dystrophy and Small for Gestational Age.

Duchenne muscular dystrophy (DMD) is a progressive recessive muscular disorder that is transmitted from mother to child. It is the most common muscular dystrophy diagnosed during childhood and early diagnosis is essential to limit its negative consequences [33]. Percy et al. [34] analysed the effectiveness of screening for female DMD carriers from four biomarkers of blood samples: serum creatine kinase (CK), haemopexin (H), pyruvate kinase (PK), and lactate dehydrogenase (LD). The available data that were analysed contain complete information on these four biomarkers of 67 women who are carriers of the progressive recessive disorder DMD and 127 women who are not carriers.

Small-for-gestational-age (SGA) infants—those with a birth weight below the 10th percentile—have been associated with increased risk of adverse perinatal outcomes [35,36]. Predicting them from the ultrasound percentile at the 35th gestational week as well as other fetal and parental variables may help to avoid these outcomes. The dataset analysed contains information of 4474 pregnancies whose births were assisted at the Miguel Servet University Hospital, between March 2012 and December 2016. The ability of the methods to predict SGA at birth was analysed from information of 7 biomarkers: ultrasound percentile at 35th gestational week, maternal age, parity, maternal body mass index (BMI), Free Beta Human Chorionic Gonatrodopin (FBHCG), Pregnacy-Associated Plasma Protein-A (PAPP-A) and paternal height.

3. Results

3.1. Simulations

The results of each method for each simulated scenario are presented as the mean of the maximum Youden indices obtained in the 1000 random samples as well as the standard de-

viation. For simplicity, we denote logistic regression, min–max approach, min–max–median approach and min–max–IQR approach by LR, MM, MMM and MMIQR, respectively.

3.1.1. Normal Distributions

Tables 1 and 2 present the results obtained for the simulated scenarios following multivariate normal distribution with different means and same means between biomarkers (in particular, $p = 4$ and $p = 10$), respectively.

Specifically, Table 1 analyses, for each sample size, biomarker dataset ($p = 4$ and $p = 10$) and for each validation method (with and without CV), the following simulated scenarios: (i) independence ($\mathbf{\Sigma_1} = \mathbf{\Sigma_2} = I$), (ii) medium correlation ($\mathbf{\Sigma_1} = \mathbf{\Sigma_2} = 0.5 \cdot I + 0.5 \cdot J$), (iii) different covariance matrices: $\mathbf{\Sigma_1} = 0.3 \cdot I + 0.7 \cdot J$ (high correlation) and $\mathbf{\Sigma_2} = 0.7 \cdot I + 0.3 \cdot J$ (low correlation) and (iv) negative correlation ($\rho = -0.1$).

Table 1. Mean and standard deviation maximum Youden indices for 1000 random samples: normal distributions. Different means between biomarkers.

Size (n_1, n_2)	Four Biomarkers				Ten Biomarkers			
	LR	**MM**	**MMM**	**MMIQR**	**LR**	**MM**	**MMM**	**MMIQR**
Independence ($\mathbf{\Sigma_1} = \mathbf{\Sigma_2} = I$)								
Resubstitution method								
(30, 30)	0.6057 (0.0957)	0.5556 (0.0962)	0.6177 (0.0884)	0.6126 (0.0877)	0.9995 (0.0055)	0.8736 (0.0563)	0.9621 (0.0338)	0.9580 (0.0356)
(100, 100)	0.5412 (0.0538)	0.4837 (0.0526)	0.5300 (0.0523)	0.5244 (0.0529)	0.9855 (0.0192)	0.8390 (0.0363)	0.9370 (0.0241)	0.9318 (0.0264)
(500, 500)	0.5120 (0.0262)	0.4447 (0.0270)	0.4808 (0.0254)	0.4752 (0.0252)	0.9589 (0.0088)	0.8155 (0.0180)	0.9173 (0.0125)	0.9163 (0.0125)
Based on CV								
(30, 30)	0.5199 (0.1101)	0.454 (0.1241)	0.4508 (0.1257)	0.4423 (0.1242)	0.8814 (0.0779)	0.7906 (0.097)	0.7791 (0.1168)	0.7574 (0.1315)
(100, 100)	0.5156 (0.0565)	0.4311 (0.0675)	0.4425 (0.0725)	0.4326 (0.0748)	0.9176 (0.0315)	0.7988 (0.0563)	0.8484 (0.0692)	0.8253 (0.0855)
(500, 500)	0.507 (0.0266)	0.4261 (0.0320)	0.4484 (0.0333)	0.4366 (0.0354)	0.9496 (0.0092)	0.8027 (0.0224)	0.8897 (0.0262)	0.8833 (0.0318)
Medium correlation ($\mathbf{\Sigma_1} = \mathbf{\Sigma_2} = 0.5 \cdot I + 0.5 \cdot J$)								
Resubstitution method								
(30, 30)	0.5480 (0.0972)	0.4610 (0.0936)	0.5062 (0.0915)	0.5074 (0.0908)	0.9796 (0.0465)	0.7344 (0.081)	0.7642 (0.0741)	0.7689 (0.0759)
(100, 100)	0.4771 (0.0577)	0.3821 (0.0545)	0.4041 (0.0537)	0.4043 (0.0536)	0.9046 (0.0327)	0.6807 (0.0471)	0.6956 (0.0454)	0.7020 (0.0450)
(500, 500)	0.4428 (0.0273)	0.3320 (0.0271)	0.3420 (0.0268)	0.3410 (0.0267)	0.8734 (0.0149)	0.6491 (0.0223)	0.6558 (0.0218)	0.6634 (0.0223)
Based on CV								
(30, 30)	0.4526 (0.1134)	0.3544 (0.1242)	0.3347 (0.1284)	0.3331 (0.1263)	0.7620 (0.0982)	0.6642 (0.1007)	0.6184 (0.1230)	0.6102 (0.1215)
(100, 100)	0.4482 (0.0627)	0.3294 (0.0689)	0.3204 (0.0733)	0.3177 (0.0737)	0.8472 (0.0347)	0.6481 (0.0544)	0.6256 (0.0748)	0.6251 (0.0704)
(500, 500)	0.4374 (0.0272)	0.3153 (0.0300)	0.3108 (0.0329)	0.3075 (0.0352)	0.8636 (0.0151)	0.6385 (0.0241)	0.6374 (0.0263)	0.6350 (0.0320)
Different correlation ($\mathbf{\Sigma_1} = 0.3 \cdot I + 0.7 \cdot J$, $\mathbf{\Sigma_2} = 0.7 \cdot I + 0.3 \cdot J$)								
Resubstitution method								
(30, 30)	0.5512 (0.0973)	0.5106 (0.0919)	0.5500 (0.0846)	0.5506 (0.0863)	0.9831 (0.0392)	0.6493 (0.0832)	0.6939 (0.0799)	0.6925 (0.0802)
(100, 100)	0.4809 (0.0559)	0.4374 (0.0546)	0.4524 (0.0529)	0.4543 (0.0531)	0.9064 (0.0336)	0.5918 (0.0504)	0.6244 (0.0494)	0.6230 (0.0495)
(500, 500)	0.4441 (0.0270)	0.3912 (0.0267)	0.3969 (0.0264)	0.3981 (0.0267)	0.8768 (0.0154)	0.5574 (0.0240)	0.5802 (0.0242)	0.5796 (0.0241)
Based on CV								
(30, 30)	0.4579 (0.1080)	0.4029 (0.1202)	0.3832 (0.1246)	0.3825 (0.1247)	0.7696 (0.0976)	0.5260 (0.1227)	0.4965 (0.1265)	0.4960 (0.1239)
(100, 100)	0.4579 (0.0602)	0.3848 (0.0691)	0.3764 (0.0741)	0.3773 (0.0723)	0.8477 (0.0331)	0.5188 (0.0805)	0.5208 (0.0789)	0.5208 (0.0809)
(500, 500)	0.4391 (0.0270)	0.3738 (0.0329)	0.3737 (0.0320)	0.3703 (0.0318)	0.8669 (0.0163)	0.5233 (0.0395)	0.5375 (0.0400)	0.5359 (0.0408)
Negative Correlation ($\rho = -0.1$)								
Resubstitution method								
(30, 30)	0.6563 (0.0914)	0.5997 (0.0893)	0.6749 (0.0835)	0.6683 (0.0830)				
(100, 100)	0.6030 (0.0527)	0.5321 (0.0534)	0.5966 (0.0501)	0.5888 (0.0512)				
(500, 500)	0.5740 (0.0239)	0.4947 (0.0251)	0.5490 (0.0242)	0.5429 (0.0259)				
Based on CV								
(30, 30)	0.5799 (0.1040)	0.4950 (0.1203)	0.5025 (0.1206)	0.4986 (0.1219)				
(100, 100)	0.5820 (0.0528)	0.4861 (0.0686)	0.5133 (0.0708)	0.4999 (0.0711)				
(500, 500)	0.5688 (0.0245)	0.4787 (0.0302)	0.5213 (0.0311)	0.5052 (0.0365)				

Table 1 shows that in the normal scenarios of independent biomarkers of different means, the MMM and MMIQR approaches show comparable performance, slightly outperforming the former, and outperforming the MM approach especially in larger samples. This behaviour is more pronounced when the number of biomarkers is larger ($p = 10$). However, they fall short of LR, which is the best performer. This behaviour is also observed in the scenario of negative correlations, although in this case the difference in performance between the summary statistics-based approaches and LR is somewhat smaller. The results show that the resubstitution method is indeed more optimistic than the one derived from cross-validation, although the conclusions on the performance comparison are broadly similar.

As for the simulated scenarios considering four biomarkers with different covariance matrices between populations, the new approaches incorporating a new summary statistic do not reach the performance of the MM approach and could be considered as approaches with similar performances. However, they outperform the MM approach, albeit only slightly when the number of biomarkers is larger ($p = 10$) and sample sizes are larger. The results show that in the scenarios of biomarkers with medium positive correlations, the new approaches do not outperform the MM approach in any of the cases based on the cross-validation method. In all cases, the LR outperforms all other methods.

Table 2 analyses the following simulated scenarios of biomarkers with the same means: (i) independence ($\mathbf{\Sigma_1} = \mathbf{\Sigma_2} = I$), (ii) medium correlation ($\mathbf{\Sigma_1} = \mathbf{\Sigma_2} = 0.5 \cdot I + 0.5 \cdot J$), (iii) different covariance matrices: $\mathbf{\Sigma_1} = 0.3 \cdot I + 0.7 \cdot J$ (high correlation) and $\mathbf{\Sigma_2} = 0.7 \cdot I + 0.3 \cdot J$ (low correlation) and (iv) different covariance matrices: $\mathbf{\Sigma_1} = 0.5 \cdot I + 0.5 \cdot J$ (medium correlation) and $\mathbf{\Sigma_2} = I$ (independents).

Table 2. Mean and standard deviation maximum Youden indices for 1000 random samples: normal distributions. Same means between biomarkers.

Size (n_1, n_2)	Four Biomarkers				Ten Biomarkers			
	LR	MM	MMM	MMIQR	LR	MM	MMM	MMIQR
Independence ($\Sigma_1 = \Sigma_2 = I$)								
Resubstitution method								
(30, 30)	0.7664 (0.082)	0.7396 (0.0803)	0.7987 (0.0702)	0.7951 (0.0709)	0.9893 (0.0319)	0.8391 (0.0642)	0.9382 (0.0436)	0.9359 (0.0439)
(100, 100)	0.7182 (0.0456)	0.6889 (0.0464)	0.7377 (0.0435)	0.7342 (0.0443)	0.9257 (0.032)	0.8013 (0.0389)	0.9051 (0.0300)	0.9037 (0.0302)
(500, 500)	0.6964 (0.0213)	0.6606 (0.0222)	0.7014 (0.0207)	0.6991 (0.0216)	0.8981 (0.0138)	0.7742 (0.0192)	0.8813 (0.0147)	0.8810 (0.0147)
Based on CV								
(30, 30)	0.6907 (0.0908)	0.6550 (0.1082)	0.6402 (0.1125)	0.6403 (0.1127)	0.7932 (0.0973)	0.7337 (0.1152)	0.7473 (0.1187)	0.7593 (0.1223)
(100, 100)	0.6977 (0.0476)	0.6432 (0.0623)	0.6607 (0.0653)	0.6526 (0.0667)	0.8667 (0.0309)	0.7377 (0.0716)	0.8222 (0.0701)	0.8127 (0.0762)
(500, 500)	0.6921 (0.0216)	0.6441 (0.0267)	0.6749 (0.0282)	0.6632 (0.0331)	0.8886 (0.0141)	0.7489 (0.0330)	0.8514 (0.0300)	0.8494 (0.0299)
Medium correlation ($\Sigma_1 = \Sigma_2 = 0.5 \cdot I + 0.5 \cdot J$)								
Resubstitution method								
(30, 30)	0.5884 (0.0990)	0.5926 (0.0919)	0.6344 (0.0876)	0.6350 (0.0870)	0.6769 (0.0949)	0.6095 (0.0882)	0.6545 (0.0842)	0.6538 (0.0838)
(100, 100)	0.5236 (0.0539)	0.5309 (0.0526)	0.5526 (0.0511)	0.5514 (0.0510)	0.5693 (0.0534)	0.5476 (0.0507)	0.5751 (0.0506)	0.5735 (0.0505)
(500,500)	0.4901 (0.0264)	0.4884 (0.0258)	0.5012 (0.0253)	0.5001 (0.0252)	0.5213 (0.0255)	0.5069 (0.0250)	0.5255 (0.0247)	0.5243 (0.0246)
Based on CV								
(30, 30)	0.4972 (0.1124)	0.4898 (0.1160)	0.4681 (0.1240)	0.4710 (0.1224)	0.4466 (0.1144)	0.4931 (0.1296)	0.4648 (0.1263)	0.4592 (0.1263)
(100, 100)	0.4945 (0.0597)	0.4777 (0.0667)	0.4729 (0.0690)	0.4714 (0.0694)	0.4974 (0.0603)	0.4798 (0.0779)	0.4724 (0.0789)	0.4668 (0.0813)
(500, 500)	0.4853 (0.0261)	0.4706 (0.0320)	0.4693 (0.0316)	0.4673 (0.0314)	0.5063 (0.0262)	0.4771 (0.0386)	0.4801 (0.0388)	0.4775 (0.0394)
Different correlation ($\Sigma_1 = 0.3 \cdot I + 0.7 \cdot J$, $\Sigma_2 = 0.7 \cdot I + 0.3 \cdot J$)								
Resubstitution method								
(30, 30)	0.5914 (0.0983)	0.6782 (0.0846)	0.7057 (0.0791)	0.7074 (0.0807)	0.6799 (0.0914)	0.7900 (0.0726)	0.8090 (0.0683)	0.8149 (0.0683)
(100, 100)	0.5304 (0.0546)	0.6211 (0.0497)	0.6325 (0.0482)	0.6348 (0.0485)	0.5769 (0.0535)	0.7425 (0.0434)	0.7502 (0.0425)	0.7559 (0.0427)
(500, 500)	0.4989 (0.0260)	0.5863 (0.0236)	0.5902 (0.0232)	0.5924 (0.0232)	0.5355 (0.0254)	0.7138 (0.0207)	0.7169 (0.0205)	0.7205 (0.0206)
Based on CV								
(30, 30)	0.5044 (0.1116)	0.5861 (0.1118)	0.5683 (0.1198)	0.5682 (0.1200)	0.4477 (0.1112)	0.7076 (0.1053)	0.6776 (0.1210)	0.6624 (0.1144)
(100, 100)	0.5028 (0.0570)	0.5771 (0.0630)	0.5722 (0.0673)	0.5711 (0.0650)	0.5060 (0.0613)	0.7007 (0.0585)	0.6982 (0.0637)	0.6770 (0.0664)
(500, 500)	0.4948 (0.0263)	0.5724 (0.0279)	0.5734 (0.0265)	0.5707 (0.0275)	0.5355 (0.0260)	0.7006 (0.0251)	0.7025 (0.0237)	0.6924 (0.0293)
Different correlation ($\Sigma_1 = 0.5 \cdot I + 0.5 \cdot J$, $\Sigma_2 = I$)								
Resubstitution method								
(30, 30)	0.6681 (0.0867)	0.7046 (0.0796)	0.7427 (0.0755)	0.7421 (0.0753)	0.7978 (0.0891)	0.8035 (0.0703)	0.8426 (0.0647)	0.8442 (0.0649)
(100, 100)	0.6181 (0.0506)	0.6548 (0.0483)	0.6796 (0.0473)	0.6795 (0.0476)	0.7308 (0.0469)	0.7601 (0.0417)	0.7919 (0.0397)	0.7927 (0.0394)
(500, 500)	0.5930 (0.0245)	0.6222 (0.0228)	0.6401 (0.0229)	0.6399 (0.0228)	0.7050 (0.0216)	0.7322 (0.0200)	0.7612 (0.0190)	0.7608 (0.0190)
Based on CV								
(30, 30)	0.5851 (0.1049)	0.6054 (0.1136)	0.5939 (0.1178)	0.5990 (0.1185)	0.5714 (0.1025)	0.6864 (0.1192)	0.6797 (0.1288)	0.6841 (0.1263)
(100, 100)	0.5940 (0.0530)	0.6076 (0.0624)	0.6144 (0.0645)	0.6167 (0.0630)	0.6742 (0.0510)	0.6996 (0.0712)	0.7113 (0.0743)	0.7081 (0.0764)
(500, 500)	0.5894 (0.0248)	0.6060 (0.0274)	0.6194 (0.0275)	0.6192 (0.0279)	0.6954 (0.0222)	0.7114 (0.0292)	0.7286 (0.0334)	0.7286 (0.0335)

Table 2 shows, as Table 1, that in simulated normal independent biomarker scenarios the MMM and MMIQR approaches outperform the MM approach, although the LR is the best performer based on CV. However, in this case of scenarios of biomarkers with same means, the difference in performance between the MMM and MMIQR approaches and the LR is significantly lower.

The results in Table 2 show that in the simulated scenarios of biomarkers with positive medium correlation, summary statistics-based methods (MM, MMM, MMIQ) could be considered to perform comparably in larger sample sizes considering the cross-validation method. Although LR still dominates over the other methods, the difference is considerably smaller than considering biomarkers with different means (Table 1).

However, the results show that in the simulated scenarios with different covariance matrices between groups, the MM approach outperforms the LR approach. Furthermore, in the scenario where biomarkers within a group are independent, the proposed MMM and MMIQR approaches slightly exceed the MM approach for larger sample sizes.

3.1.2. Non-Normal Distributions

Table 3 shows the results obtained from the following simulated scenarios following a log-normal distribution: (i) biomarkers with different means and medium correlation ($\mathbf{\Sigma_1} = \mathbf{\Sigma_2} = 0.5 \cdot I + 0.5 \cdot J$), (ii) biomarkers with different means and different covariance matrices: $\mathbf{\Sigma_1} = 0.3 \cdot I + 0.7 \cdot J$ (high correlation) and $\mathbf{\Sigma_2} = 0.7 \cdot I + 0.3 \cdot J$ (low correlation), (iii) biomarkers with same means and medium correlation ($\mathbf{\Sigma_1} = \mathbf{\Sigma_2} = 0.5 \cdot I + 0.5 \cdot J$), (iv) biomarkers with same means and different covariance matrices: $\mathbf{\Sigma_1} = 0.3 \cdot I + 0.7 \cdot J$ (high correlation) and $\mathbf{\Sigma_2} = 0.7 \cdot I + 0.3 \cdot J$ (low correlation), and (v) biomarkers with same means and different covariance matrices: $\mathbf{\Sigma_1} = 0.5 \cdot I + 0.5 \cdot J$ (medium correlation) and $\mathbf{\Sigma_2} = I$ (independents).

Table 3. Mean and standard deviation maximum Youden indices for 1000 random samples: Non-normal distributions. Log-normal.

Size (n_1, n_2)	Four Biomarkers				Ten Biomarkers			
	LR	MM	MMM	MMIQR	LR	MM	MMM	MMIQR
Different means. Medium correlation ($\Sigma_1 = \Sigma_2 = 0.5 \cdot I + 0.5 \cdot J$)								
Resubstitution method								
(30, 30)	0.5290 (0.1009)	0.4506 (0.0988)	0.5053 (0.0914)	0.5076 (0.0915)	0.9650 (0.0593)	0.6969 (0.0853)	0.7593 (0.0752)	0.7624 (0.0755)
(100, 100)	0.4594 (0.0588)	0.3798 (0.0558)	0.4036 (0.0537)	0.4052 (0.0538)	0.8765 (0.0378)	0.6571 (0.0486)	0.6905 (0.0459)	0.6948 (0.0458)
(500, 500)	0.4278 (0.0283)	0.3323 (0.0270)	0.3425 (0.0269)	0.3428 (0.0269)	0.8428 (0.0168)	0.6319 (0.0231)	0.6513 (0.0219)	0.6552 (0.0221)
Based on CV								
(30, 30)	0.4272 (0.1230)	0.3709 (0.1132)	0.3344 (0.1241)	0.3367 (0.1273)	0.7254 (0.1006)	0.6652 (0.1006)	0.5902 (0.1203)	0.5948 (0.1203)
(100, 100)	0.4277 (0.0634)	0.3348 (0.0660)	0.3148 (0.0681)	0.3172 (0.0675)	0.8161 (0.0403)	0.6420 (0.0556)	0.5875 (0.0826)	0.5947 (0.0776)
(500, 500)	0.4220 (0.0282)	0.3113 (0.0316)	0.3054 (0.0342)	0.3051 (0.0340)	0.8320 (0.0176)	0.6289 (0.0235)	0.6029 (0.0452)	0.6074 (0.0419)
Different means. Different correlation ($\Sigma_1 = 0.3 \cdot I + 0.7 \cdot J$, $\Sigma_2 = 0.7 \cdot I + 0.3 \cdot J$)								
Resubstitution method								
(30, 30)	0.5139 (0.1018)	0.5047 (0.0915)	0.5423 (0.0853)	0.5400 (0.0860)	0.9614 (0.0577)	0.6431 (0.0845)	0.6901 (0.0803)	0.6892 (0.0798)
(100, 100)	0.4476 (0.0609)	0.4318 (0.0544)	0.4468 (0.0534)	0.4470 (0.0534)	0.8637 (0.0407)	0.5847 (0.0506)	0.6214 (0.0492)	0.6206 (0.0493)
(500, 500)	0.4157 (0.0286)	0.3881 (0.0271)	0.3934 (0.0269)	0.3936 (0.0270)	0.8304 (0.0176)	0.5498 (0.0238)	0.5770 (0.0242)	0.5764 (0.0242)
Based on CV								
(30, 30)	0.4118 (0.1166)	0.4063 (0.1183)	0.3872 (0.1236)	0.3925 (0.1243)	0.7080 (0.1006)	0.5195 (0.1189)	0.5258 (0.1184)	0.5329 (0.1180)
(100, 100)	0.4161 (0.0628)	0.3889 (0.0670)	0.3760 (0.0712)	0.3852 (0.0686)	0.8001 (0.0409)	0.5044 (0.0831)	0.5308 (0.0747)	0.5379 (0.0742)
(500, 500)	0.4097 (0.0287)	0.3761 (0.0294)	0.3726 (0.0318)	0.3751 (0.0303)	0.8196 (0.0182)	0.5195 (0.0402)	0.5314 (0.0436)	0.5412 (0.0385)
Same means. Medium correlation ($\Sigma_1 = \Sigma_2 = 0.5 \cdot I + 0.5 \cdot J$)								
Resubstitution method								
(30, 30)	0.5729 (0.1016)	0.5911 (0.0924)	0.6325 (0.0887)	0.6331 (0.0881)	0.6800 (0.1042)	0.6058 (0.0900)	0.6527 (0.0851)	0.6523 (0.0843)
(100, 100)	0.5103 (0.0549)	0.5292 (0.0522)	0.5520 (0.0514)	0.5514 (0.0514)	0.5538 (0.0549)	0.5448 (0.0510)	0.5736 (0.0505)	0.5733 (0.0505)
(500, 500)	0.4811 (0.0265)	0.4873 (0.0258)	0.5002 (0.0253)	0.4995 (0.0253)	0.5120 (0.0260)	0.5056 (0.0252)	0.5251 (0.0248)	0.5242 (0.0247)
Based on CV								
(30, 30)	0.4867 (0.1111)	0.4975 (0.1120)	0.4767 (0.1181)	0.4803 (0.1180)	0.4316 (0.1182)	0.4926 (0.1162)	0.4866 (0.1182)	0.4910 (0.1208)
(100, 100)	0.4858 (0.0600)	0.4693 (0.0667)	0.4687 (0.0674)	0.4717 (0.0684)	0.4866 (0.0613)	0.4669 (0.0766)	0.4798 (0.0701)	0.4827 (0.0714)
(500, 500)	0.4779 (0.0266)	0.4585 (0.0353)	0.4626 (0.0334)	0.4652 (0.0333)	0.4997 (0.0263)	0.4631 (0.0453)	0.4800 (0.0421)	0.4824 (0.0399)
Same means. Different correlation ($\Sigma_1 = 0.3 \cdot I + 0.7 \cdot J$, $\Sigma_2 = 0.7 \cdot I + 0.3 \cdot J$)								
Resubstitution method								
(30, 30)	0.5445 (0.1042)	0.6730 (0.0849)	0.6990 (0.0802)	0.6979 (0.0813)	0.6437 (0.0957)	0.7783 (0.0739)	0.7988 (0.0708)	0.7993 (0.0712)
(100, 100)	0.4714 (0.0559)	0.6172 (0.0499)	0.6286 (0.0486)	0.6291 (0.0486)	0.5129 (0.0548)	0.7292 (0.0446)	0.7382 (0.0440)	0.7393 (0.0442)
(500, 500)	0.4365 (0.0265)	0.5840 (0.0237)	0.5880 (0.0235)	0.5885 (0.0235)	0.4590 (0.0259)	0.7010 (0.0206)	0.7050 (0.0205)	0.7057 (0.0205)
Based on CV								
(30, 30)	0.4525 (0.1149)	0.5914 (0.1127)	0.5753 (0.1167)	0.5866 (0.1154)	0.4020 (0.1141)	0.7084 (0.1008)	0.6903 (0.1066)	0.7007 (0.1006)
(100, 100)	0.4472 (0.0576)	0.5854 (0.0589)	0.5744 (0.0630)	0.5819 (0.0596)	0.4386 (0.0621)	0.6977 (0.0558)	0.6904 (0.0592)	0.6949 (0.0567)
(500, 500)	0.4324 (0.0271)	0.5744 (0.0257)	0.5723 (0.0264)	0.5743 (0.0254)	0.4435 (0.0264)	0.6917 (0.0229)	0.6905 (0.0234)	0.6916 (0.0234)
Same means. Different correlation ($\Sigma_1 = 0.5 \cdot I + 0.5 \cdot J$, $\Sigma_2 = I$)								
Resubstitution method								
(30, 30)	0.5932 (0.0974)	0.6942 (0.0825)	0.7296 (0.0787)	0.7287 (0.0789)	0.7222 (0.0960)	0.7924 (0.0738)	0.8340 (0.0669)	0.8334 (0.0669)
(100, 100)	0.5248 (0.0543)	0.6477 (0.0503)	0.6712 (0.0484)	0.6708 (0.0486)	0.6206 (0.0524)	0.7506 (0.0423)	0.7850 (0.0408)	0.7844 (0.0406)
(500, 500)	0.4966 (0.0268)	0.6188 (0.0238)	0.6350 (0.0231)	0.6347 (0.0231)	0.5770 (0.0250)	0.7276 (0.0206)	0.7560 (0.0195)	0.7556 (0.0195)
Based on CV								
(30, 30)	0.5074 (0.1089)	0.6123 (0.1107)	0.6031 (0.1127)	0.6187 (0.1085)	0.4840 (0.1158)	0.7215 (0.1004)	0.7031 (0.1135)	0.7261 (0.1033)
(100, 100)	0.5006 (0.0566)	0.6182 (0.0557)	0.6138 (0.0625)	0.6256 (0.0577)	0.5474 (0.0565)	0.7286 (0.0508)	0.7178 (0.0687)	0.7365 (0.0565)
(500, 500)	0.4918 (0.0270)	0.6100 (0.0249)	0.6141 (0.0293)	0.6198 (0.0255)	0.5619 (0.0258)	0.7221 (0.0217)	0.7310 (0.0304)	0.7399 (0.0237)

The results in Table 3 show that, under cross-validation, in the simulated scenarios of biomarkers with different means and medium correlation, the MMM and MMIQR approaches do not outperform the MM approach and the LR significantly dominates over

the rest. In scenarios with different covariance matrices between groups with a higher number of biomarkers ($p = 10$), although LR remains the best performer, the MMM and MMIQR approaches outperform the MM approach, with the MMIQR approach being slightly superior.

However, in the biomarker scenarios with the same means, considering medium correlation between biomarkers, the results show that for larger sample sizes the MMM and MMIQR approaches outperform the MM approach in this case, with the MMIQR approach being very slightly superior.

For the simulated scenarios of biomarkers with the same means and different covariance matrices between groups, the summary statistics-based approaches outperform logistic regression. For the scenario with independent biomarkers in one group, the MMIQR approach outperforms the others by applying the cross-validation method.

3.1.3. High-Dimensional Scenarios: Normal Distributions

Table 4 analyses the following simulated scenarios of $p = 100$ normal biomarkers with the same means for a smaller sample size ($n_1 = n_2 = 30$): (i) independence ($\mathbf{\Sigma_1} = \mathbf{\Sigma_2} = I$) and (ii) different covariance matrices: $\mathbf{\Sigma_1} = 0.5 \cdot I + 0.5 \cdot J$ (medium correlation) and $\mathbf{\Sigma_2} = I$ (independents).

Table 4. Mean and standard deviation maximum Youden indices for 1000 random samples. High-dimensional scenarios ($p = 100$): Normal distributions. Same means between biomarkers.

Independence ($\mathbf{\Sigma_1} = \mathbf{\Sigma_2} = I$)				
Size (n_1, n_2)	**LR**	**MM**	**MMM**	**MMIQR**
Resubstitution method				
(30, 30)	1 (0)	0.4364 (0.094)	0.7358 (0.077)	0.7307 (0.0785)
Based on CV				
(30, 30)	0.1684 (0.0952)	0.2602 (0.1084)	0.4577 (0.1372)	0.4527 (0.1391)
Different correlation ($\mathbf{\Sigma_1} = 0.5 \cdot I + 0.5 \cdot J$, $\mathbf{\Sigma_2} = I$)				
Size (n_1, n_2)	**LR**	**MM**	**MMM**	**MMIQR**
Resubstitution method				
(30, 30)	1 (0)	0.9121 (0.0518)	0.9277 (0.047)	0.9308 (0.0459)
Based on CV				
(30, 30)	0.1568 (0.0912)	0.8433 (0.0718)	0.807 (0.1226)	0.7067 (0.1477)

Table 4 shows that in these scenarios where the size of the variables is larger than the sample size, the LR model is significantly overfitted. The results obtained based on CV show that statistics-based algorithms are superior to LR. This difference is most noticeable in the scenario with different covariance matrices. In this scenario, our approaches (MMM and MMIQR) fail to outperform the MM approach. However, in the scenario of independent biomarkers, our approaches outperform it. These results are in line with the conclusions drawn from the results in Section 3.1.1.

3.2. Computational Times

Our proposed approach uses the information provided by a set of predictive markers in three new markers derived from them. This is an automated calculus that it is not time-consuming. From them, we estimated the coefficients of the model by an extensive search using a step by step approach. We want to remark that computational times for this approach are reasonably independent of the sample size. In our simulations, the computational times of our proposed approaches ranged from 8 to 154 seconds using a sample

size ranging from 60 to 20,000 individuals showing that our algorithm is computationally tractable in high dimensionality.

3.3. Real Datasets

3.3.1. Duchenne Muscular Dystrophy Dataset

Figure 1 represents the information of each biomarker for each group (carrier and non-carrier groups).

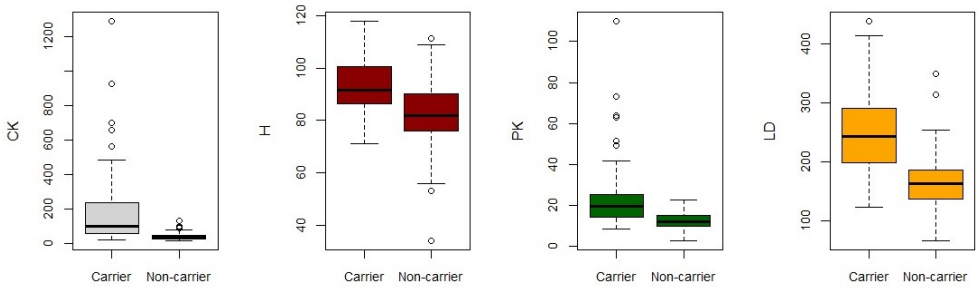


Figure 1. Information from DMD dataset.

As the variances of the four biomarkers differ from each other, it was necessary to normalise the values of each biomarker before applying the methods based on summary statistics. This normalisation is important to avoid different units of measurement and to ensure the consistent use of these approaches.

The analysed methods were applied to this dataset with the aim of finding the optimal combination of the four biomarkers that achieves the maximum Youden index. The estimates of the Youden index of each biomarker (CK, H, PK, LD) in a univariate way were 0.6124, 0.4172, 0.5079, and 0.5776.

Table 5 shows the maximum Youden index achieved as well as their respective sensitivity and specificity values, using both the resubstitution and the cross-validation-based methods. The MMM approach outperforms the MM approach, although it does not reach the performance of the LR, which is the best performer, achieving a Youden index of 0.7948 after cross-validation.

Table 5. Maximum Youden index for each method. DMD dataset.

Methods	Resubstitution Method			Based on CV		
	Youden	Sensitivity	Specificity	Youden	Sensitivity	Specificity
LR	0.8106	0.8657	0.9449	0.7948	0.8657	0.9291
MM	0.7335	0.8358	0.8976	0.5602	0.8358	0.7244
MMM	0.8019	0.8806	0.9213	0.6472	0.6866	0.9606
MMIQR	0.7948	0.8657	0.9291	0.5535	0.6716	0.8819

The results rounded to four decimal places are displayed.

3.3.2. Small for Gestational Age Dataset

Figure 2 represents the information of each biomarker for each group (SGA and non-SGA group). Before applying the summary statistics-based methods, a normalisation was applied.

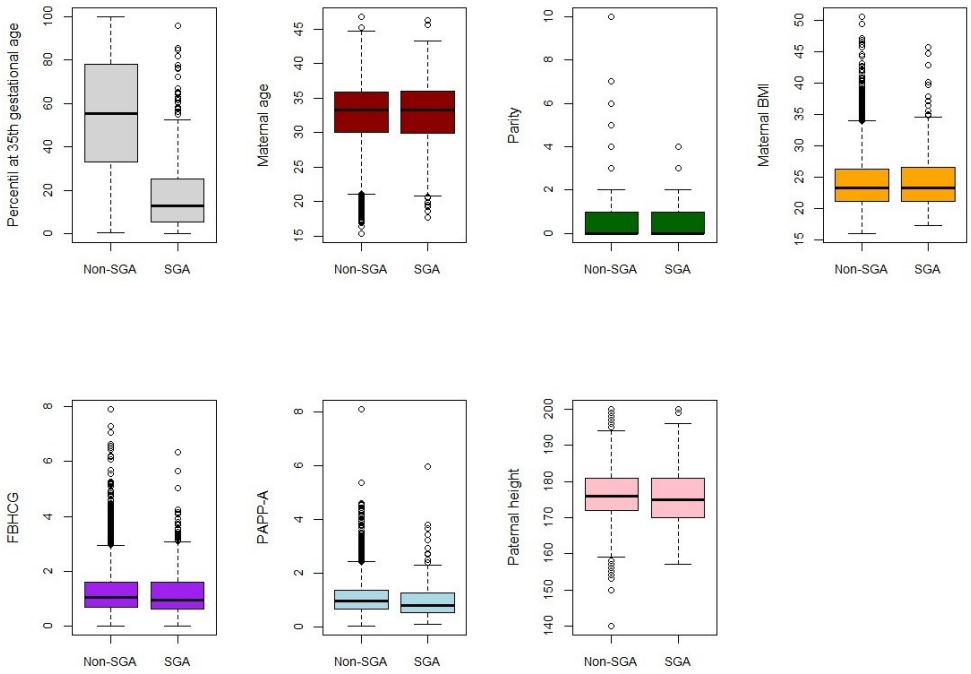


Figure 2. Information from SGA dataset.

Table 6 shows the maximum Youden index achieved for all methods studied as well as their respective sensitivity and specificity values, using both the resubstitution and the cross-validation-based methods. The MMM and MMIQR approaches outperform the MM approach, although they do not match the performance of the LR, which is the best performer, achieving a Youden index of 0.5901 after cross-validation. In this example, the MMM and MMIQR approaches outperform the MM approach more markedly than the previous example, which includes a smaller number of biomarkers ($p = 4$). These results therefore show that for a larger number of biomarkers, the benefit of using the approaches, incorporating a new summary statistic (MMM, MMIQR vs. MM approach), was greater. These conclusions were also drawn from the simulated data. Specifically, this is a scenario with a predictor, the estimated percentile weight at the 35th week, with a greater discrimination ability than the rest of predictor variables. It seems that in this case, the MM approach cannot capture the added discrimination ability of the set of biomarkers; in this case, the improvement of MMM and MMIQ approaches is more clear.

Table 6. Maximum Youden index for each method. SGA dataset.

Methods	Resubstitution method			Based on CV		
	Youden	Sensitivity	Specificity	Youden	Sensitivity	Specificity
LR	0.5971	0.7639	0.8332	0.5901	0.7687	0.8214
MM	0.1631	0.9108	0.2523	0.1422	0.6554	0.4868
MMM	0.3692	0.6988	0.6704	0.3400	0.7373	0.6026
MMIQR	0.3644	0.7373	0.6270	0.2322	0.6410	0.5913

Although for none of the real examples analysed did the summary statistics-based approaches outperform LR, it is clear from this analysis that the proposed approaches outperform the MM approach, which has been shown in the simulation to have superior discrimination ability than the LR for specific scenarios of different covariance matrices between groups and the same univariate predictive ability of biomarkers.

4. Discussion and Conclusions

In this study, we propose new summary statistics-based approaches and compare their performances with the min–max approach and logistic regression under Youden index optimisation, which are computationally tractable approaches regardless of the problem dimension. The methods were applied to both simulated data and two real datasets.

The results of the comparison using cross-validation (Tables 1–3) suggest that a logistic regression should be used over the other summary statistics-based approaches when the number of biomarkers is smaller than the sample size, except in scenarios of biomarkers with the same predictive ability and different covariance matrices between groups (Table 2), where summary statistics-based approaches are recommended, which outperformed logistic regression. In fact, generally, differences in performance were smaller when considering scenarios with biomarkers with the same means (Table 2). The study by Kang et al. [20], but under AUC optimisation, led to a similar conclusion regarding optimal scenarios for the min–max approach.

The results showed that the proposed new approaches outperform the min–max approach in general when independent biomarkers are included. This difference increases when the number of biomarkers is larger. These results may be expected as independent information as well as a larger number of biomarkers may contribute to the fact that it is insufficient to consider only the minimum and maximum values of biomarkers. Although no significant differences were found between the MMM and MMIQR approaches, the MMIQR approach slightly outperformed the MMM approach in non-normal distributions and the MMM approach overall slightly outperformed the others in all other scenarios. This could indicate that in asymmetric distributions, the incorporation of a dispersion measure could be a good option.

In scenarios where the number of biomarkers is higher than the sample size (Table 4), the results showed that the logistic regression overfitted the data. Our approaches outperformed the min–max approach when biomarkers were independent. These results are in line with previous results and demonstrate the advantage of our approaches which, in addition to always being computationally feasible, do not overfit the data due to their inherent feature of dimensionality reduction involving all biomarkers across the summary information.

The performance of the analysed methods achieved on the real datasets showed similar conclusions to those deduced from the simulation. Although in none of the real examples analysed did the summary statistics-based approaches outperform the logistic approach, in both cases the proposed methods outperformed the MM approach.

The application of dimensionality reduction techniques before applying a method could be an alternative to avoid overfitting. The selection of biomarkers with high predictive power might seem to be a solution for dimensionality reduction. However, it has been shown that the predictive power alone of each biomarker may not be the best approach for selection. Pinsky and Zhou [37] demonstrated that in fact statistical correlation patterns had a greater implication on the combination of higher performing biomarkers. Therefore, a high number of biomarkers with low predictive ability could be available, yet their combination achieves higher performance. Therefore, methods that use information from a large number of biomarkers, while being computationally tractable, are necessary.

Our proposed distribution-free approach incorporates information from all the original biomarkers through their summary statistics, while the problem remains computationally tractable ($p = 3$). Pepe and Thomson [15] proposed non-parametric approaches to optimise linear models under AUC optimisation but with reduced dimensionality. Ma and Huang [38] and Wang et al. [39] introduced semiparametric approaches that approximated the empirical AUC by a sigmoid function. Komori et al. [40] proposed a method based on a boosting algorithm for maximisation of the AUC. They used cross-validation techniques to select the best model. None of those methods is fully non-parametric, and their performance relies on a good selection of some parameters.

Liu et al. [21] proposed a non-parametric approach that linearly combines the minimum and maximum values of the biomarkers, involving only a single coefficient search. Therefore,

the problem is computable as it is reduced to estimating a single parameter regardless of the number of biomarkers. In addition to being able to take advantage of the information from multiple biomarkers that are measured with the same unit, another advantage of this min–max procedure is that repeated measurements of a single biomarker can be combined as two new markers and can therefore also be an option for longitudinal studies. However, using the information from the minimum and maximum biomarker values may not be sufficient in terms of discrimination, especially when the number of biomarkers is high. The proposed algorithm incorporates more information from the original variables than the min–max approach (three variables). A stepwise algorithm is performed to estimate the linear combination. All this implies that our proposal always remains computationally tractable ($p = 3$) and has inherent dimensionality reduction properties.

With regard to the strengths of our study, we can list the following. These algorithms are fully non-parametric approaches and do not require distribution assumption for their application. They use information from all biomarkers, incorporate more information than the min–max approach and are also computationally tractable regardless of the number of biomarkers. In addition, we verified their superiority in some simulated scenarios, highlighting that they are good alternatives in predictive models.

However, our study has some limitations. In the comparison analysis we used a unique standard method, the logistic regression. We did not perform a comparison with other methodologies as kernel approaches or machine learning algorithms; therefore, we cannot propose our algorithm as the best method in any scenario.

For future work, a more extensive study comparing more methods could be carried out. Applying dimensionality reduction techniques (such as principal component analysis) before applying the methods could also be explored. Extending the comparison study for a more extensive simulation, e.g., taking into account scenarios with larger sample sizes, is also proposed as a line of future work. It is also proposed to analyse the new methods proposed in other real datasets. An example could be the one analysed by Liu et al. (autism disease), where conditions were optimal for the MM approach, and it could thus be explored as to whether in this case the proposed approaches outperform the MM approach. Readers are also encouraged to adapt our proposed approaches using a different linear combination estimation (a different stepwise algorithm) as well as to explore their performances considering other target metrics.

In summary, in this article, we present to the scientific community a new approach combining summary statistics of biomarkers (min–max–median, min–max–IQR approaches) that has the advantage of being always computable and not being subject to any distributional assumptions. This approach aims to discriminate between two groups (disease and non-disease) and can therefore be a key to and helpful in clinical decision making. The Youden index, which is a widely used metric in the absence of consensus, was considered as an objective function. Comparisons in various scenarios allowed us to discover the scenarios in which the approaches are most optimal. Specifically:

- Our proposed approaches outperform classical logistic regression using original biomarkers when the number of biomarkers is larger than the sample size.
- Additionally, they performed well in scenarios of biomarkers with the same predictive ability and different covariance matrices between groups.
- Our approaches performed better than the min–max approach when biomarkers are independent and in the real scenarios.

Author Contributions: Conceptualisation, R.A.-G. and L.M.E.; methodology, R.A.-G., L.M.E. and G.S.; software, R.A.-G. and L.M.E. and R.d.-H.-A.; validation, R.A.-G. and L.M.E.; formal analysis, R.A.-G. and L.M.E.; investigation, R.A.-G., L.M.E. and G.S.; resources, R.A.-G., L.M.E. and R.S.-C.; data curation, R.A.-G., L.M.E. and R.S.-C.; writing—original draft preparation, R.A.-G. and L.M.E.; writing—review and editing, R.A.-G., L.M.E., G.S., R.d.-H.-A. and R.S.-C.; visualisation, R.A.-G. and L.M.E.; supervision, R.A.-G., L.M.E., G.S. and R.d.-H.-A.; funding acquisition, G.S. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded by project MTM2017-83812-P of MINECO.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Actual data are from published studies.

Data Availability Statement: Data from the SGA dataset were retrieved from the Miguel Servet University Hospital database and are not shared. The Duchenne Muscular Dystrophy dataset can be found at https://hbiostat.org/data/, accessed on 30 August 2021.

Acknowledgments: The research of Luis M. Esteban and Gerardo Sanz was supported by the Stochastic Models Research group of the Goverment of Aragon. The research of Rocío Aznar-Gimeno and Rafael del-Hoyo-Alonso was supported by the Integration and Development of Big Data and Electrical Systems (IODIDE) group.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AUC	Area Under the ROC curve
BMI	Body Mass Index
CK	Serum Creatine Kinase
CV	Cross-Validation
FBHCG	Free Beta Human Chorionic Gonatrodopin
H	Haemopexin
IQR	Interquartile Range
LD	Lactate Dehydrogenase
LR	Logistic Regression
MM	Min–Max approach
MMIQR	Min–Max–IQR approach
MMM	Min–Max–Median approach
PAPP-A	Pregnacy-Associated Plasma Protein-A
pAUC	Partial Area Under the ROC curve
PK	Pyruvate Kinase
ROC	Receiver Operating Characteristic
SGA	Small-for-Gestational-Age

References

1. Amini, M.; Kazemnejad, A.; Zayeri, F.; Amirian, A.; Kariman, N. Application of adjusted-receiver operating characteristic curve analysis in combination of biomarkers for early detection of gestational diabetes mellitus. *Koomesh* **2019**, *21*, 751–758.
2. Yu, S. A Covariate-Adjusted Classification Model for Multiple Biomarkers in Disease Screening and Diagnosis. Ph.D. Thesis, Kansas State University, Manhattan, AR, USA, 2019.
3. Bansal, A.; Pepe, M.S. When does combining markers improve classification performance and what are implications for practice? *Stat. Med.* **2013**, *32*, 1877–1892. [CrossRef] [PubMed]
4. Mi, G.; Li, W.; Nguyen, T.S. Characterize and Dichotomize a Continuous Biomarker. In *Statistical Methods in Biomarker and Early Clinical Development*; Fang, L., Su, C., Eds.; Springer: Cham, Switzerland, 2019; pp. 23–38. [CrossRef]
5. Esteban, L.M.; Sanz, G.; Borque, A. Linear combination of biomarkers to improve diagnostic accuracy in prostate cancer. *Monografías Matemáticas García de Galdeano* **2013**, *38*, 35–84.
6. Youden, W.J. Index for rating diagnostic tests. *Cancer J.* **1950**, *3*, 32–35. [CrossRef]
7. Lyu, T.; Ying, Z.; Zhang, H. A new semiparametric transformation approach to disease diagnosis with multiple biomarkers. *Stat. Med.* **2019**, *38*, 1386–1398. [CrossRef] [PubMed]
8. Ma, H.; Yang, J.; Xu, S.; Liu, C.; Zhang, Q. Combination of multiple functional markers to improve diagnostic accuracy. *J. Appl. Stat.* **2020**, 1–20. [CrossRef]
9. Ahmadian, R.; Ercan, I.; Sigirli, D.; Yildiz, A. Combining binary and continuous biomarkers by maximizing the area under the receiver operating characteristic curve. *Commun. Stat. Simul. Comput.* **2020**, 1–14. [CrossRef]
10. Su, J.Q.; Liu, J.S. Linear combinations of multiple diagnostic markers. *J. Am. Stat. Assoc.* **1993**, *88*, 1350–1355. [CrossRef]
11. Yan, L.; Tian, L.; Liu, S. Combining large number of weak biomarkers based on AUC. *Stat. Med.* **2015**, *34*, 3811–3830. [CrossRef]
12. Xu, T.; Fang, Y.; Rong, A.; Wang, J. Flexible combination of multiple diagnostic biomarkers to improve diagnostic accuracy. *BMC Med. Res. Methodol.* **2015**, *15*, 1–7. [CrossRef]
13. Nigmatullin, R.R. The statistics of the fractional moments: Is there any chance to "read quantitatively" any randomness? *Signal Process.* **2006**, *86*, 2529–2547. [CrossRef]

14. Nigmatullin, R.R.; Lino, P.; Maione, G. *New Digital Signal Processing Methods*; Springer Publishing: New York, NY, USA, 2020.
15. Pepe, M.S.; Thompson, M.L. Combining diagnostic test results to increase accuracy. *Biostatistics* **2000**, *1*, 123–140. [CrossRef] [PubMed]
16. Pepe, M.S.; Cai, T.; Longton, G. Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics* **2006**, *62*, 221–229. [CrossRef] [PubMed]
17. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36. [CrossRef]
18. Esteban, L.M.; Sanz, G.; Borque, A. A step-by-step algorithm for combining diagnostic tests. *J. Appl. Stat.* **2011**, *38*, 899–911. [CrossRef]
19. Kang, L.; Xiong, C.; Crane, P.; Tian, L. Linear combinations of biomarkers to improve diagnostic accuracy with three ordinal diagnostic categories. *Stat. Med.* **2013**, *32*, 631–643. [CrossRef]
20. Kang, L.; Liu, A.; Tian, L. Linear combination methods to improve diagnostic/prognostic accuracy on future observations. *Stat. Methods Med. Res.* **2016**, *25*, 1359–1380. [CrossRef]
21. Liu, C.; Liu, A.; Halabi, S. A min–max combination of biomarkers to improve diagnostic accuracy. *Stat. Med.* **2011**, *30*, 2005–2014. [CrossRef]
22. Yin, J.; Tian, L. Optimal linear combinations of multiple diagnostic biomarkers based on Youden index. *Stat. Med.* **2014**, *33*, 1426–1440. [CrossRef]
23. Liu, A.; Schisterman, E.F.; Zhu, Y. On linear combinations of biomarkers to improve diagnostic accuracy. *Stat. Med.* **2005**, *24*, 37–47. [CrossRef]
24. Yin, J.; Tian, L. Joint inference about sensitivity and specificity at the optimal cut-off point associated with Youden index. *Comput. Stat. Data Anal.* **2014**, *77*, 1–13. [CrossRef]
25. Yin, J.; Tian, L. Joint confidence region estimation for area under ROC curve and Youden index. *Stat. Med.* **2014**, *33*, 985–1000. [CrossRef]
26. Ma, H.; Halabi, S.; Liu, A. On the use of min-max combination of biomarkers to maximize the partial area under the ROC curve. *J. Probab. Stat.* **2019**, 8953530. [CrossRef]
27. Yu, W.; Park, T. Two simple algorithms on linear combination of multiple biomarkers to maximize partial area under the ROC curve. *Comput. Stat. Data Anal.* **2015**, *88*, 15–27. [CrossRef]
28. Yan, Q.; Bantis, L.E.; Stanford, J.L.; Feng, Z. Combining multiple biomarkers linearly to maximize the partial area under the ROC curve. *Stat. Med.* **2018**, *37*, 627–642. [CrossRef]
29. Perkins, N.J.; Schisterman, E.F. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *Am. J. Epidemiol.* **2006**, *163*, 670–675. [CrossRef]
30. Friedman, J.; Hastie, T.; Tibshirani, R. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer: New York, NY, USA, 2009; pp. 219–259.
31. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020. Available online: http://www.r-project.org/index.html (accessed on 20 September 2021).
32. Walker, S.H.; Duncan, D.B. Estimation of the probability of an event as a function of several independent variables. *Biometrika* **1967**, *54*, 167–179. [CrossRef]
33. Mohamed, K.; Appleton, R.; Nicolaides, P. Delayed diagnosis of Duchenne muscular dystrophy. *Eur. J. Paediatr. Neurol.* **2000**, *4*, 219–223. [CrossRef]
34. Percy, M.E.; Andrews, D.F.; Thompson, M.W. Duchenne muscular dystrophy carrier detection using logistic discrimination: Serum creatine kinase, hemopexin, pyruvate kinase, and lactate dehydrogenase in combination. *Am. J. Med. Genet. A* **1982**, *13*, 27–38. [CrossRef]
35. Savirón-Cornudella, R.; Esteban, L.M.; Aznar-Gimeno, R.; Dieste-Pérez, P.; Pérez-López, F.R.; Campillos, J.M.; Castán-Larraz, B.; Sanz, G.; Tajada-Duaso, M. Prediction of Late-Onset Small for Gestational Age and Fetal Growth Restriction by Fetal Biometry at 35 Weeks and Impact of Ultrasound–Delivery Interval: Comparison of Six Fetal Growth Standards. *J. Clin. Med.* **2021**, *10*, 2984. [CrossRef]
36. Savirón-Cornudella, R.; Esteban, L.M.; Tajada-Duaso, M.; Castán-Mateo, S.; Dieste-Pérez, P.; Cotaina-Gracia, L.; Lerma-Puertas, D.; Sanz, G.; Pérez-López, F.R. Detection of Adverse Perinatal Outcomes at Term Delivery Using Ultrasound Estimated Percentile Weight at 35 Weeks of Gestation: Comparison of Five Fetal Growth Standards. *Fetal Diagn. Ther.* **2020**, *47*, 104–114. [CrossRef] [PubMed]
37. Pinsky, P.F.; Zhu, C.S. Building multi-marker algorithms for disease prediction—The role of correlations among markers. *Biomark. Insights* **2011**, *6*, BMI-S7513. [CrossRef] [PubMed]
38. Ma, S.; Huang, J. Combining multiple markers for classification using ROC. *Biometrics* **2007**, *63*, 751–757. [CrossRef] [PubMed]
39. Wang, Z.; Chang, Y.; Ying, Z.; Zhu, L.; Yang, Y. A parsimonious threshold-independent protein feature selection method through the area under receiver operating characteristic curve. *Bioinformatics* **2007**, *23*, 2788–2794. [CrossRef]
40. Komori, O.; Eguchi, S. A boosting method for maximizing the partial area under the ROC curve. *BMC Bioinform.* **2010**, *11*, 314. [CrossRef]

6.2. Comparación del enfoque MMM con modelos de ML (XGBoost)

En esta sección se presenta el artículo [7], cuyo desarrollo complementa el estudio presentado en la sección anterior.

En el estudio anterior [8], a pesar de llevar a cabo una comparación exhaustiva en diversos escenarios simulados y obtener resultados que podrían ofrecer ciertas pautas sobre cuándo sería apropiado el uso del algoritmo, la cantidad de enfoques comparados podría ser insuficiente. Aunque la regresión logística es un método de comparación estándar, para poder verificar mejor la adecuación de nuestros enfoques, era importante compararlos con otros algoritmos del estado del arte que han sido desarrollados más recientemente.

Los algoritmos de aprendizaje automático (Machine Learning, ML), además de la regresión logística, son también métodos eficaces para abordar problemas de clasificación binaria. En los últimos años, su aplicación en investigación clínica ha aumentado, gracias a sus resultados prometedores y a su facilidad de implementación con una carga computacional asequible. En concreto, el algoritmo XGBoost (presentado en la sección 4.3.2) es uno de los algoritmos que más se han utilizado, superando en rendimientos a otros enfoques de ML y convencionales en diversas aplicaciones. A diferencia de las limitaciones de los modelos de combinación lineal, el algoritmo XGBoost es capaz de capturar relaciones no lineales en los datos. Sin embargo, no es inherentemente interpretable como lo son los modelos lineales. A pesar de esto, en los últimos años se ha trabajado en el concepto de inteligencia artificial explicable (XAI, siglas en inglés), que busca ofrecer interpretabilidad y transparencia a través de técnicas aplicadas a modelos que no son inherentemente explicables.

Estos resultados impulsaron nuestra investigación. En concreto, nuestro objetivo fue extender la comparación del rendimiento de los enfoques propuestos (MMM/IQR), incluyendo también el algoritmo de ML XGBoost. Consideramos una diversidad de escenarios simulados que incluyeron distribuciones simétricas y asimétricas, diferente tamaño, así como diferentes capacidades discriminatorias y correlaciones de los biomarcadores. Para cada escenario, se entrenó cada método utilizando muestras aleatorias de la distribución subyacente. Los parámetros estimados y el punto de corte óptimo seleccionado bajo el criterio del índice de Youden se aplicaron a la muestra de validación. Los índices de Youden obtenidos en la muestra de validación se utilizaron para evaluar el rendimiento de los enfoques. Además, se evaluaron los métodos en dos conjuntos de datos reales (riesgo de mortalidad materna y distrofia muscular de Duchenne), utilizando un procedimiento de validación cruzada.

Los resultados muestran que los enfoques de ML superan en rendimiento a nuestros enfoques en escenarios con biomarcadores con diferentes capacidades predictivas. Sin embargo, nuestros enfoques siguen obteniendo mejor rendimiento en escenarios de biomarcadores con la misma capacidad predictiva y diferentes matrices de covarianza entre grupos. Esto refuerza las conclusiones extraídas del trabajo previo, evidenciando que para estos escenarios, nuestros enfoques siguen siendo más óptimos que otros algoritmos del estado del arte, como el XGBoost.

A continuación se presenta el artículo, resultado de este trabajo.

symmetry

MDPI

Article

Comparing the Min–Max–Median/IQR Approach with the Min–Max Approach, Logistic Regression and XGBoost, Maximising the Youden Index

Rocío Aznar-Gimeno [1,*], Luis M. Esteban [2,*], Gerardo Sanz [3] and Rafael del-Hoyo-Alonso [1]

[1] Department of Big Data and Cognitive Systems, Instituto Tecnológico de Aragón (ITAINNOVA), 50018 Zaragoza, Spain
[2] Department of Applied Mathematics, Escuela Universitaria Politécnica de La Almunia, Universidad de Zaragoza, La Almunia de Doña Godina, 50100 Zaragoza, Spain
[3] Department of Statistical Methods, Institute for Biocomputation, Physics of Complex Systems-BIFI, University of Zaragoza, 50009 Zaragoza, Spain
* Correspondence: raznar@itainnova.es (R.A.-G.); lmeste@unizar.es (L.M.E.)

Abstract: Although linearly combining multiple variables can provide adequate diagnostic performance, certain algorithms have the limitation of being computationally demanding when the number of variables is sufficiently high. Liu et al. proposed the min–max approach that linearly combines the minimum and maximum values of biomarkers, which is computationally tractable and has been shown to be optimal in certain scenarios. We developed the Min–Max–Median/IQR algorithm under Youden index optimisation which, although more computationally intensive, is still approachable and includes more information. The aim of this work is to compare the performance of these algorithms with well-known Machine Learning algorithms, namely logistic regression and XGBoost, which have proven to be efficient in various fields of applications, particularly in the health sector. This comparison is performed on a wide range of different scenarios of simulated symmetric or asymmetric data, as well as on real clinical diagnosis data sets. The results provide useful information for binary classification problems of better algorithms in terms of performance depending on the scenario.

Keywords: classification; linear combination; Youden index; min–max approach; min–max–median approach; min–max-IQR approach; logistic regression; XGBoost

Citation: Aznar-Gimeno, R.; Esteban, L.M.; Sanz, G.; del-Hoyo-Alonso, R. Comparing the Min–Max–Median/IQR Approach with the Min–Max Approach, Logistic Regression and XGBoost, Maximising the Youden Index. *Symmetry* **2023**, *15*, 756. https://doi.org/10.3390/sym15030756

Academic Editor: Juan Luis García Guirao

Received: 31 January 2023
Revised: 12 March 2023
Accepted: 15 March 2023
Published: 19 March 2023

1. Introduction

The linear combination of multiple biomarkers is often used in clinical practice [1] for disease diagnosis due to its ease of interpretation and performance [2], which is usually superior to considering each biomarker separately [3–8]. These new biomarkers are key for disease screening or understanding the evolution of a disease after diagnosis. As an example, Prostate-specific antigen (PSA) is the most used biomarker to diagnose prostate cancer, although it lacks the necessary sensitivity and specificity. The prostate health index (PHI) and 4Kscore are new biomarkers with greater predictive ability derived from linear models that include PSA [9].

To assess diagnostic accuracy, statistics derived from the receiver operating characteristic (ROC) curve, such as the area under the ROC curve (AUC) [10] or the Youden index [11], are often used. In this context, the development of binary classification model approaches that maximise the AUC has been extensively studied in the literature. Many of these studies have been the basis for the formulation of subsequently published improved approaches under ROC-curve-derived optimality criteria.

Su and Liu [12] formulated the optimal linear model that maximises the AUC under the assumption of multivariate normality. This normality assumption is often not easy to observe in real clinical practice, being too demanding in part due to the symmetry that

biomarkers must meet. For many diseases, the progression or advanced stages of them are associated with high values of the diagnostic tests, so these types of variables tend to follow asymmetric distributions. Results from a prostate cancer screening cohort show a clear asymmetry of PSA in the Canadian population [13]. This limitation was solved by Pepe et al. [14,15], who proposed a distribution-free approach for the estimation of the linear model that maximises AUC based on the Mann–Whitney U-statistic [16]. This approach is based on discrete optimisation under extensive search on the parameter vector of biomarkers coefficients. Although the statistical foundation underpinning the approach proposed by Pepe et al. has been the basis for subsequent approaches, it has the drawback of being computationally infeasible when the number of biomarkers is greater than or equal to three. To address this computational limitation, Pepe et al. [14,15] suggested the use of stepwise algorithms based on selecting and estimating, at each step, the best linear combination of two biomarkers, including at each step a new biomarker. This proposal for partial optimisations at each step was later implemented by Esteban et al. [17] and Kang et al. [18]. Esteban et al. included tie-handling strategies, and Kang et al. proposed a simpler and less demanding approach by setting the order of biomarker inclusion at the beginning of the algorithm. Liu et al. [19] proposed an approach, called the min–max approach, which is computationally tractable regardless of the number of biomarkers. This is because it is based on the linear combination of the minimum and maximum values of biomarkers under the optimisation of the Mann–Whitney U-statistic of the AUC, involving the search for a single optimal coefficient. Despite its computational advantage, it has been shown to generally achieve lower accuracy than other approaches that use information from all biomarkers, such as stepwise approaches, but shows superiority in some scenarios [3,4,18].

In diagnostic or binary classification problems where combinations of continuous biomarkers are estimated, dichotomisation of the resulting continuous value, i.e., establishing a cut-off point, is often key, as it provides a classification rule that allows this classification of patients into groups [20]. In this sense, the Youden index is a good criterion for choosing the best cut-off point to dichotomise a biomarker [21] and is an appropriate summary of the performance of the diagnostic model [22]. For example, the Youden index takes a cut-off value of 45.9 for PHI in nonfused biopsies [23]. The Youden index maximises the sum of sensitivity and specificity, giving equal weight to both metrics, so that it can be considered as the symmetrical point that maximises both metrics simultaneously.

Therefore, although there are different metrics that provide the optimal cut-off point, in the absence of consensus, with no clear reason to optimise either sensitivity or specificity, the Youden index provides that optimal balance, being the most used parameter to choose a threshold.

Although the area under the ROC curve is the most studied diagnostic assessment statistic in the literature, other statistics such as the Youden index are also used in different clinical studies and provide accurate categorisation. The algorithms under AUC optimality cited above were used as a basis for the formulation of subsequent approaches under Youden index maximisation. Based on the stepwise approach of Kang et al. [18], Yin and Tian [24] conducted a study under Youden index optimisation. Aznar-Gimeno et al. [25] developed the stepwise algorithm suggested by Pepe et al. [14,15] under Youden index maximisation and compared its performance with other approaches in the literature, modified under Youden index maximisation, such as Yin and Tian's stepwise approach [24], the min–max approach [19], logistic regression [26], a parametric method with multivariate normality and a non-parametric kernel smoothing method. Although Aznar-Gimeno et al. demonstrated that their proposed approach achieved acceptable performance, superior in some scenarios, it has the computational limitation of being difficult to approach when the number of biomarkers increases. The min–max approach, which solves this computational problem through the linear combination of the minimum and maximum values, did not prove to be sufficient in terms of discrimination, except in a few specific scenarios.

Maintaining the advantage of not being subject to any distributional assumptions, being computationally tractable regardless of the number of original biomarkers while incorporating more information through a new summary statistic (the median or the interquartile range), Aznar-Gimeno et al. [27] proposed the so-called min–max–median and min–max-IQR approaches. These approaches are based on estimating the linear combination of these three variables using the proposed stepwise algorithm [25]. Aznar-Gimeno et al. compared the proposed algorithms with the min–max algorithm and logistic regression. The aim was to compare computationally tractable methods, regardless of the number of biomarkers. In this sense, cancer shows a substantial clinical heterogeneity, and the max–min derived approach tries to capture the potential variation underlying the biological heterogeneity.

Machine learning algorithms have been increasingly used in various fields of application [28] and, in particular, in clinical practice and medical research [29–34], due to their performance potential and efficiency. There are different machine learning and deep learning techniques that have been applied in the area of health from different sources of information covering different formats ranging from numerical data to text or images [35]. Numerous studies have applied and evaluated these techniques in recent years with different objectives in the healthcare domain [36], such as predicting events, diagnosing or prognosing diseases or cancers [37–40]. Analysing their association with patient biomarkers such as demographic data, clinical data, pharmacology, genetics, medical imaging or wearable sensors, [41] (among others), is a challenge that needs to be addressed in a way that prevents or detects the disease early.

Deep learning has been used to assist in the identification of genes and associated proteomics and metabolomics profiles to detect cancers at early stages [42–44]. Concerning the early detection of breast cancer, Mahesh et al. [45] evaluated the Naive Bayes classifier, the Decision Tree classifier, Random Forest and their ensembles. Botlagunta et al. [46] assessed nine machine learning methods for breast cancer metastasis classification, including logistic regression, k-nearest neighbours, decision trees, random forest, gradient boosting, and eXtreme Gradient Boosting (XGBoost) [47]. Rustam et al. [48] compared the performance of Support Vector Machine (SVM) and Naive Bayes for prostate cancer patient classification. Huo et al. [49] also evaluated the effectiveness of machine learning models for prostate cancer prediction, including SVM, decision tree, random forest, XGBoost, and adaptive boosting (Adaboost). Sabbagh et al. [50] applied logistic regression and XGBoost techniques to the prediction of lymph node metastasis in prostate cancer patients using clinicopathologic features. Khan et al. [51] propose a self-normalised multiview convolutional neural network model with adaptive boosting (AdaBoost-SNMV-CNN) for lung cancer nodule detection in computed tomography scans. Regarding diabetes, Saheb-Honar et al. [52] examined the classification ability of logistic regression, decision tree, and random forest in identifying the relationship between type 2 diabetes and its risk factors. Budholiya et al. [53] present a diagnostic system that employs an optimised XGBoost classifier with the aim of predicting the occurrence of heart disease. Ensemble models, combining machine learning and deep learning approaches, provide personalized patient treatment strategies based on medical histories and diagnostics [54]. The versatility of deep learning models is clear, with applications for omics data types, as well as histopathology-based genomic inference, providing perspectives on the integration of different data types to develop decision support tools [55], but few of them have yet demonstrated real-world medical utility [56].

The primary drawback of one of these algorithms compared to techniques based on linear models is the lack of explainability and interpretability of the models. One of the key reasons why these tools may not be effectively implemented and integrated into routine clinical practice is due to the lack of transparency and explainability of the models. Explainable artificial intelligence (XAI) is attracting much interest in medicine [57] and, fortunately, in recent years, work has been carried out on the concept of XAI, which provides techniques that also offer explainability and transparency of these models.

XGBoost [47] is one of the most widely used machine learning algorithms of recent times. This is due to its ease of implementation and good results, proving to be a leader in many competitions and state-of-the-art studies [58]. The XGBoost algorithm assumes no normality and combines several weak prediction models, which are usually decision trees, improving its predictivity and accuracy. This type of model shows versatility as it depends on some parameters relating to the building trees that can be optimized.

In terms of machine learning algorithms, our work focuses on analysing the predictive capacity of logistic regression and XGBoost. Numerous studies have compared the performance of logistic regression and XGBoost in the health domain in recent years [59–66]. Unlike logistic regression or other statistical approaches based on linear models, XGBoost allows capturing non-linear relationships, one of the main reasons for its popularity. However, although XGBoost is an effective tool in healthcare and has been in demand in recent years, demonstrating good performance, it does not always outperform conventional statistical methods such as logistic regression [62,66,67]. The choice of the optimal model will depend on the problem and the type of data. Therefore, it is always necessary to conduct a comprehensive comparative study to analyse the performance of algorithms in different scenarios in order to obtain useful information and establish certain guidelines.

Due to the enormous number of data available nowadays by the advances in technology, it has been shown that it is essential to develop non-parametric biomarker combination models that are computationally tractable, regardless of the number of initial biomarkers. In this sense, our proposed approaches (min–max–median/IQR approach) reduce the dimensional problem by capturing the heterogeneity of the information through summary statistics. Although studies comparing the performance of different machine learning techniques have increased in the literature in recent years, so far, there are no studies comparing the performance of our proposed approaches with machine learning models such as XGBoost, which has been in high demand in recent years and which can capture more complex relationships than the statistical linear methods compared in other studies [25,27]. The aim of our work was to compare the performance of our proposed min–max–median/IQR approaches with the min–max approach and the machine learning algorithms known as logistic regression and XGBoost, maximising the Youden index. For this purpose, they were compared on a wide range of simulated symmetric or asymmetric data scenarios, as well as on real clinical diagnostic datasets.

We provide a novel approach based on three main basic characteristics of the set of predictor variables, the maximum, minimum and median or IQR to capture the larger discrimination ability to summarize in these three parameters. On the other hand, from a different perspective, we train and validate additive tree models trying to capture the sum of the predictive ability of all predictor variables. The results of this work provide the reader with useful information that can serve as a guide for the choice of the most suitable algorithm for binary classification problems depending on the characteristics and behaviour of the data.

2. Materials and Methods

This section introduces some notations and the non-parametric approach of Pepe et al. [14,15], which forms the basis for our min–max–median/IQR approaches. In the following, we explain our proposed approaches (min–max–median/IQR) and the algorithms with which we compare performance: min–max approach, logistic regression and XGBoost. These algorithms were adapted by optimising the Youden index. Finally, the simulated scenarios and real datasets are detailed, as well as the validation procedure. The entire study was conducted using the free software R (The R Foundation for statistical computing, Vienna, Austria) [68]. The code of the whole study can be found in Supplementary Material.

2.1. Background

Consider the following binary classification problem where p is the number of biomarkers, n_1 is the number of case individuals (with disease) and n_2 is the number of control

individuals (healthy individuals). If X_{kij} denotes the value of the j^{th} variable or biomarker ($j = 1, \ldots, p$) for the i^{th} individual of group k=1,2 (disease and non-disease), then $\mathbf{X_{ki}}$ is the vector of biomarkers for the i^{th} individual of group k = 1,2 and $\mathbf{X_1} = (\mathbf{X_{11}}, \ldots, \mathbf{X_{1n_1}})$ and $\mathbf{X_2} = (\mathbf{X_{21}}, \ldots, \mathbf{X_{2n_2}})$. Therefore, the linear combination of each group is expressed as $\mathbf{Y_k} = \boldsymbol{\beta}^T \mathbf{X_k}$, $k = 1, 2$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ denotes the parameter vector.

Defined in the above notation, by definition, the Youden index (J) of the linear combination is expressed as:

$$\begin{aligned} J &= max_c\{Sensitivity(c) + Specificity(c) - 1\} \\ &= max_c\{F_{\mathbf{Y_2}}(c) - F_{\mathbf{Y_1}}(c)\} \end{aligned} \tag{1}$$

where c denotes the cut-off point and $F_{\mathbf{Y_k}}(c) = P(\mathbf{Y_k} \leq c)$ the cumulative distribution function of random variable $\mathbf{Y_k}$. Denoting by $c_{\boldsymbol{\beta}} = \{c : max_c(F_{\mathbf{Y_2}}(c) - F_{\mathbf{Y_1}}(c))\}$ the optimal cut-off point, the expression of the empirical estimate of the Youden index is:

$$\begin{aligned} \hat{J}_{\boldsymbol{\beta}} &= \hat{F}_{\mathbf{Y_2}}(\hat{c}_{\boldsymbol{\beta}}) - \hat{F}_{\mathbf{Y_1}}(\hat{c}_{\boldsymbol{\beta}}) \\ &= \frac{\sum_{i=1}^{n_2} I(\boldsymbol{\beta}^T \mathbf{X_{2i}} \leq \hat{c}_{\boldsymbol{\beta}})}{n_2} - \frac{\sum_{i=1}^{n_1} I(\boldsymbol{\beta}^T \mathbf{X_{1i}} \leq \hat{c}_{\boldsymbol{\beta}})}{n_1} \end{aligned} \tag{2}$$

where I denotes the indicator function.

Pepe et al.'s Approach

Pepe and Thompson [14] proposed a distribution-free approach (without any distribution assumptions) to estimate the linear model that maximizes the AUC based on the Mann–Whitney U-statistic [16]. The basis on which their proposed approach lies is mainly in the property of invariance of the ROC curve to any monotonic transformation.

Specifically, Pepe and Thompson propose the following linear model:

$$L_\beta(\mathbf{X}) = X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{3}$$

where p denotes the number of biomarkers, X_i the biomarker $i \in [1, \ldots, p]$ and β_i the parameter to be estimated. Observe that they did not include an intercept in the linear model (3), and the coefficient associated with the first variable X_1 is 1. This is because the ROC curves for $L_\beta(\mathbf{X})$ (3) and $L_\alpha(\mathbf{X}) = \alpha_0 + \alpha_1 L_\beta(\mathbf{X}), \alpha_1 > 0$ are the same, so it is enough to consider (3). Thus, considering the optimal parameter vector, the maximum empirical AUC based on the Mann–Whitney U statistic would be given by the following expression:

$$\widehat{AUC} = \frac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I(L_\beta(\mathbf{X_{1i}}) > L_\beta(\mathbf{X_{2j}})) + \frac{1}{2} I(L_\beta(\mathbf{X_{1i}}) = L_\beta(\mathbf{X_{2j}}))}{n_1 \cdot n_2}$$

Note that searching the entire possible parameter vector space $\mathbb{R}^{p-1}$ and possible coefficient-variable combinations is computationally intractable. To overcome this limitation, Pepe et al. suggested estimating the parameter vector through a discrete optimisation over 201 equally spaced values between -1 and 1. This is because selecting β in $[-1, 1]$ is equivalent to covering the range $(-\infty, \infty)$ since the AUC of $X_i + \beta X_j$ for $\beta > 1$ and $\beta < -1$ is the same as $\alpha X_i + X_j$ for $\alpha = \frac{1}{\beta} \in [-1, 1]$. Even so, this optimisation is computationally costly for dimensions $p \geq 3$. To address this, Pepe et al. [14,15] suggested the use of stepwise algorithms, in which a new variable is included in each step, selecting the best combination of two variables. In this way, the problem is transformed into a computationally tractable problem by estimating a single parameter $p - 1$ times using a linear combination of two variables.

Both the model formulation and the empirical search are the basis for the formulation of the min–max approach and our proposed algorithms (min–max–median/IQR) that extend the min–max approach, which are explained below.

2.2. Min–Max Approach

Liu et al. [19] proposed the so-called min–max approach (**MM**), which is a distribution-free approach, as proposed by Pepe et al. [14,15], but with the advantage of being computationally tractable, regardless of the number of original biomarkers. The idea of this approach is to calculate the minimum and maximum values of the p biomarkers and to consider the optimal linear combination of these two markers, involving the search for a single optimal coefficient. Specifically, the original aim is to estimate the β parameter such as the combination

$$X_{min} + \beta X_{max} \tag{4}$$

which maximizes AUC based on the Mann–Whitney U statistic, where X_{min} and X_{max} are the minimum and maximum values of the original p biomarkers for each individual, respectively.

Considering the Youden index as our target metric to maximise, the min–max approach can be adapted by selecting the optimal parameter β and cut-off point c_β that maximises the following expression

$$\hat{J}_\beta = \frac{\sum_{i=1}^{n_2} I\left(X_{2i,max} + \beta X_{2i,min} \leq \hat{c}_\beta\right)}{n_2} - \frac{\sum_{i=1}^{n_1} I\left(X_{1i,max} + \beta X_{1i,min} \leq \hat{c}_\beta\right)}{n_1} \tag{5}$$

where $X_{ki,max} = max_{1\leq j\leq p}\left(X_{kij}\right)$ and $X_{ki,min} = min_{1\leq j\leq p}\left(X_{kij}\right)$ for $k = 1, 2$ and each $i = 1, \ldots, n_k$, and $\beta \in [-1, 1]$, following Pepe et al's. suggestion of the empirical search of β.

The procedure can be summarised as follows:

1. For each i individual, the biomarkers with minimum and maximum values (X_{min} and X_{max}) are considered as the new 2 markers (for simplicity, X_1 and X_2).
2. For each of the 201 possible values of β, the value of the linear combination ($X_1 + \beta X_2$) is calculated for each i individual and the optimal cut-off point is chosen, i.e., the one that maximises the Youden index.
3. The linear combination that achieves the highest Youden index is the optimal combination.

2.3. Min–Max–Median/IQR Approach

Aznar-Gimeno et al. proposed new non-parametric approaches, so-called min–max–median (**MMM**) and min–max-IQR (**MMIQR**) [27], which extend the idea of the min–max approach by applying our proposed stepwise algorithm [25], following the suggestion of Pepe et al. [14,15]. The aim was to include more information in the model while remaining computationally affordable, although more intensive.

Specifically, the idea behind the approaches is to reduce the dimension of the problem by reducing the number of original p biomarkers to three, considering the summary statistic information of the original variables, i.e., the minimum, maximum, median, or interquartile range (IQR). Our approach extends the min–max approach as it incorporates a new summary statistic, turning the problem into a three-variable linear combination optimisation problem. As suggested by Pepe et al., a stepwise algorithm that we developed is used in this case, where the best linear combination of two variables is selected, including a new variable in each step.

Below, we provide a detailed description of the procedure for the min–max–median approach (note that the min–max-IQR approach follows the same steps).

1. Firstly, for each i individual, the minimum, maximum, and median values of p biomarkers are calculated:

$$X_{ki,max} = \max_{1\leq j\leq p}\left(X_{kij}\right), X_{ki,min} = \min_{1\leq j\leq p}\left(X_{kij}\right), X_{ki,median} = median_{1\leq j\leq p}\left(X_{kij}\right) \tag{6}$$

where $k = 1, 2$ and $i = 1, \ldots, n_k$. These values are considered as the three new variables (X_1, X_2 and X_3, for simplicity). Specifically, from now on, the problem is to estimate

the optimal linear combination of these three variables using the proposed stepwise algorithm.

2. The first step of the stepwise approach is to choose the combination(s) of the two variables that maximises the Youden index such that

$$\hat{J}_{\beta_2} = \frac{\sum_{i=1}^{n_2} I\left(X_{2ij} + \beta_2 X_{2ik} \leq \hat{c}_{\beta_2}\right)}{n_2} - \frac{\sum_{i=1}^{n_1} I\left(X_{1ij} + \beta_2 X_{1ik} \leq \hat{c}_{\beta_2}\right)}{n_1} \quad \beta_2 \in [-1,1], \quad \forall j \neq k = 1,\ldots,p \tag{7}$$

using empirical search proposed by Pepe et al. In other words, for each variable pair, for each value of the 201 (β values), the linear combination is calculated and the optimal cut-off point that maximises the Youden index is selected. That linear combination for which the optimal cut-off point has obtained the maximum Youden index is chosen in this step. Suppose, for simplicity, the optimal linear combination $X_{ki1} + \beta_2 X_{ki2}$.

3. The last step is to include the remaining variable (X_3) and select the optimal linear combination(s). Specifically, the previously chosen linear combination ($X_{ki1} + \beta_2 X_{ki2}$) is considered as a new variable and the idea of the previous point (2) is re-applied. Therefore, either combination (8) or (9) that maximizes the Youden index is chosen as the final optimal combination of the linear model.

$$\hat{J}_{\beta_3} = \frac{\sum_{i=1}^{n_2} I\left((X_{2i1} + \beta_2 X_{2i2}) + \beta_3 X_{2i3} \leq \hat{c}_{\beta_3}\right)}{n_2} - \frac{\sum_{i=1}^{n_1} I\left((X_{1i1} + \beta_2 X_{1i2}) + \beta_3 X_{1i3} \leq \hat{c}_{\beta_3}\right)}{n_1} \quad \beta_3 \in [-1,1] \tag{8}$$

$$\hat{J}_{\beta_3} = \frac{\sum_{i=1}^{n_2} I\left(\beta_3(X_{2i1} + \beta_2 X_{2i2}) + X_{2i3} \leq \hat{c}_{\beta_3}\right)}{n_2} - \frac{\sum_{i=1}^{n_1} I\left(\beta_3(X_{1i1} + \beta_2 X_{1i2}) + X_{1i3} \leq \hat{c}_{\beta_3}\right)}{n_1} \quad \beta_3 \in [-1,1] \tag{9}$$

For ease, a single optimal linear combination is considered in steps 2 and 3. However, the maximum Youden index can be reached for different linear combinations. Our algorithm considers all ties, which can be broken in the last stage (step 3) or not.

Our proposed approaches are openly available to the scientific community through the R library `SLModels` [69]. The library also incorporates the min–max algorithm adapted for the optimisation of the Youden index (previous section).

2.4. Logistic Regression

The logistic regression (**LR**) (or logit regression) [26] is a statistical model that provides the probability of an observation/individual i belonging to an output category, given its set of independent variables $\mathbf{X}_i$, through the logistics function:

$$\begin{aligned} P(\mathbf{Y_i} = 1|\mathbf{X_i}) &= \frac{1}{1 + e^{-\boldsymbol{\beta}^T \mathbf{X_i}}} = \frac{e^{\boldsymbol{\beta}^T \mathbf{X_i}}}{1 + e^{\boldsymbol{\beta}^T \mathbf{X_i}}} \\ \log \frac{P(\mathbf{Y_i} = 1|\mathbf{X_i})}{1 - P(\mathbf{Y_i} = 1|\mathbf{X_i})} &= \boldsymbol{\beta}^T \mathbf{X_i} \end{aligned} \tag{10}$$

where β is the vector of parameters to estimate by means of the maximum likelihood method.

2.5. Extreme Gradient Boosting (XGBoost)

XGBoost (eXtreme Gradient Boosting, **XGB**) is a scalable tree boosting system that was developed by Chen and Guestrin [47]. It is a specific optimised implementation of gradient boosting and is therefore based on the principle of sequential order ensemble learning, where errors are minimized (loss function) using a gradient descent algorithm. Specifically, XGBoost is a decision tree ensemble based on the idea of training several weak learners (base learners) sequentially in order to create a strong learner with higher accuracy. During training, the parameters of each weak model are adjusted by minimising the objective function, and each new model is trained to correct the errors of the previous ones. Correctly

and incorrectly predicted results receive different scores that are finally weighted to obtain a final result.

The XGBoost algorithm, unlike those presented above, has both parameters and hyperparameters. Hyperparameters are values of model settings that must be set during the training process to control the behaviour and performance of the model.

Considering the XGBoost algorithm as an ensemble base learners of decision trees, the loss function at iteration t to minimise has the following expression:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y_i}^{(t-1)} + f_t(\mathbf{X}_i)\right) + \Omega(f_t) \tag{11}$$

where l is the loss term and $\Omega(f) = \gamma T + \frac{1}{2}\lambda||\omega||^2$ is the regularisation term, which penalizes the complexity of the model, avoiding over-fitting. y_i indicates the real output, $\hat{y_i}^{(t-1)}$ the prediction of the i^{th} individual at the $(t-1)^{th}$ iterations, f denotes the base learners, T the number of leaves of the tree and ω the weights of the leaves. γ represents the minimum loss reductions needed to split a leaf node of the tree. The larger γ is, the more conservative the algorithm will be.

The complexity of the model can also be limited through the maximum-depth hyperparameter, which specifies the maximum number of levels of the tree, where each level represents a division of the data based on a variable. Another possible regularisation hyperparameter is shrinkage, which reduces the step size to make the boosting process more conservative. In other words, it decreases the influence of each individual tree and allows future trees to improve the model. Random subsampling is another regularisation technique that can be used. In the case of a column subsample, the hyperparameter specifies the subsample fraction of columns to be used to construct each tree. The same idea is for rows, where, if the value is less than 1, a random subset of rows (observations/individuals) is selected for each tree.

The XGBoost model was applied using the free software R library `xgboost`. Specifically, in this study, the following hyperparameters were adjusted over a set of possibilities:

- `nrounds`: Number of decision trees in the final model.
- `gamma` (γ): Minimum loss reduction required to split a node.
- `eta` (shrinkage, learning rate): Step size shrinkage.
- `max_depth`: Maximum depth of the tree.
- `colsample_bytree`: Subsample ratio of columns.
- `subsample`: Subsample ratio of the training instances.

Table 1 shows the hyperparameter possibilities space explored in the study. The explored values of maximum tree depth for datasets with fewer variables were lower than those with higher dimensions. For the selection of the best combination of hyperparameters, the grid search technique was used, and 5-fold cross-validation was performed on the training set. Finally, the model was trained on the entire training dataset with the selected optimal hyperparameters. The early stopping technique was used as an additional technique to avoid over-fitting by stopping the training if there was no improvement in 10 iterations in a row.

Table 1. Search space of the hyperparameters explored.

`nrounds`	`gamma`	`eta`	`max_depth`	`colsample_bytree`	`subsample`
50,100,200	0,0.5	0.1,0.3	[2,20]	0.5,1	0.5,1

2.6. Simulations

A wide range of simulated data were explored in order to analyse and compare the performance of the algorithms previously discussed. Specifically, scenarios simulating different biomarker distributions, discrimination capabilities, and correlation between them were analysed, considering $p = 4$ and $p = 10$ biomarkers, and smaller ($n_1 = n_2$= 50)

and larger ($n_1 = n_2$= 500) sample sizes. As for the biomarker distributions, both symmetric distributions (normal distributions) and asymmetric distributions (different marginal distributions and multivariate log-normal skewed distribution) were simulated.

The scenarios with different marginal distributions were simulated with 4 biomarkers following chi-square, normal, gamma and exponential distributions via normal copula with a dependence parameter between biomarkers of 0.7 for the case population (patients; diseased population) and 0.3 for the control population (healthy; non-diseased population). More specifically, the biomarkers for the control population were considered to be marginally distributed as $\chi^2_{0.1}$, $N(0.1)$, $\Gamma(0.1)$, $Exp(0.1)$ and $\chi^2_{0.1}$, $N(0.6)$, $\Gamma(0.8)$, $Exp(0.1)$ for case population. Scenarios under log-normal distribution were generated from the configurations of the simulated scenarios under a normal distribution and then exponentiated.

Concerning the scenarios of normal distributions, the null vector $m_2 = \vec{0}$ was considered as the mean vector of the non-diseased population. With respect to the mean vector of the diseased population (m_1), scenarios with the same means$m_1 = (1.0, 1.0, \dots)^T$, i.e., the same predictive ability, and different means $m_1 = (0.2, 0.5, 1.0, 0.7)^T$, $m_1 = (0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0)^T$, were explored. For simplicity, the variance of each biomarker was set to 1, so that covariances are equivalent to correlations. The same correlation value was considered for all pairs of biomarkers. Let Σ_1 and Σ_2 be the variance–covariances matrices for diseased and non-diseased populations, respectively. The following scenarios with different biomarker means were analysed:

- Independents ($\Sigma_1 = \Sigma_2 = I$).
- High correlation ($\Sigma_1 = \Sigma_2 = 0.3 \cdot I + 0.7 \cdot J$).
- Different correlation between groups ($\Sigma_1 = 0.3 \cdot I + 0.7 \cdot J, \Sigma_2 = 0.7 \cdot I + 0.3 \cdot J$).
- Negative correlation ($\rho = -0.1$).

where I denotes the identity matrix and J the all-one matrix. Regarding scenarios with the same biomarker means, the following were explored:

- Low correlation ($\Sigma_1 = \Sigma_2 = 0.7 \cdot I + 0.3 \cdot J$).
- Different correlation between groups ($\Sigma_1 = 0.3 \cdot I + 0.7 \cdot J, \Sigma_2 = 0.7 \cdot I + 0.3 \cdot J$).
- Different correlation between groups with biomarkers independents in the non-diseased population ($\Sigma_1 = 0.5 \cdot I + 0.5 \cdot J, \Sigma_2 = I$).

2.7. Application in Real Datasets

The methods being examined were also applied in two real clinical datasets: for the diagnosis of Duchenne muscular dystrophy and for maternal mortality risk.

Duchenne muscular dystrophy (DMD) is a genetic disorder passed down from a mother to her children, causing progressive muscle weakness and wasting. Percy et al. [70] analysed the effectiveness of detecting this disorder using four biomarkers extracted from blood samples: serum creatine kinase (CK), haemopexin (H), pyruvate kinase (PK) and lactate dehydrogenase (LD). The dataset was obtained at https://hbiostat.org/data/, accessed on 30 January 2023. After removing observations with missing data, the dataset used contains information on the four biomarkers of 67 women who are carriers of the progressive recessive disorder DMD and 127 women who are not carriers.

Maternal mortality refers to the death of a woman due to a pregnancy-related cause. It is one of the main concerns of the Sustainable Development Goals (SDG) of the United Nations. The dataset used for analysing maternal mortality was obtained at [71] (Maternal Health Risk), which contains information on the following six risk factors for maternal mortality: age in years during pregnant, upper value of blood pressure in mmHg (SystolicBP), lower value of blood pressure in mmHg (DiastolicBP), blood glucose levels in mmol/L (BS), body temperature in ºF (BodyTemp) and a normal resting heart rate in beats per minute (HeartRate). An IoT-based risk monitoring system was used to gather this information from various hospitals, community clinics, and maternal healthcare centres in the rural areas of Bangladesh. The level of risk intensity was also provided by differentiating three categories: low (406 women), medium (336 women) and high (272 women). To adapt data to our study, the following binary problems were considered: (i) predicting high or medium

risk versus low risk and (ii) predicting high risk versus medium or low risk. The original dataset contains repeated data, outliers and anomalous data that may be due to errors in data retrieval. In our study, data from women aged 13–50 years were considered, duplicate rows were removed, and two observations were removed with heart rate values of 7 beats per minute, which is an erroneous value. Finally, the dataset used in the study contained 196 low-risk, 86 medium-risk and 95 high-risk observations.

2.8. Validation

A total of 59 simulated data scenarios were explored, considering different sample sizes, number of biomarkers, distributions, discriminatory ability, and correlations. For each simulated scenario, each method was trained considering random samples from the underlying distribution (100) and validated using new data simulated with the same configuration (100). For real data, a 10-fold cross-validation procedure was performed.

During training, the model parameters were estimated, and the optimal cut-off point that maximises the Youden index was obtained. The estimated model and the cut-off point selected were applied to the validation set. The Youden indices obtained in the validation set are shown in the tables in the following sections.

3. Results

This section presents the results of the performance achieved by each of the methods studied, both for the simulated scenarios and for the real data datasets. We denote the logistic regression, XGBoost, min–max approach, min–max–median approach, and min–max-IQR approach by LR, XG, MM, MMM, and MMIQR, respectively.

3.1. Simulations

The results obtained from the 100 random samples for each scenario are presented as the mean of the maximum Youden indices as well as the standard deviation. The following sections present the results for the symmetric (normal) and non-symmetric distribution scenarios.

3.1.1. Symmetric Distributions

Tables 2 and 3 display the results obtained from the simulated scenarios for $p = 4$ biomarkers following a multivariate normal distribution with different means and the same means, respectively. The code can be found at Supplementary Material: Chapter 1, Sections A.1.–A.5.

The results in Table 2 show, in general, a superiority of logistic regression over the other algorithms, except in the scenario of different correlations and larger sample sizes, where XGBoost significantly outperforms it. Our proposed algorithms (MMM/MMIQR) show similar performance to the min–max approach or superior, particularly when biomarkers are independent or negatively correlated.

Table 3 presents the results for biomarkers with the same predictive ability. The conclusions derived are different from those in the Table 2 above (different mean). Logistic regression achieves the highest average performance value in the same and low correlation scenarios, although the rest of the algorithms obtained very close values. The summary-statistics-based methods (MM, MMM, MMIQR) and especially our proposed algorithms (MMM/MMIQR) outperformed the other algorithms in scenarios with different correlations. XGBoost outperforms logistic regression in scenarios with different correlations and large sample sizes.

Table 2. Normal distributions. Different means. Four biomarkers.

Size (n_1,n_2)	LR	XG	MM	MMM	MMIQR
Independents					
(50,50)	**0.453 (0.0919)**	0.3878 (0.1016)	0.3696 (0.1037)	0.3884 (0.1075)	0.3938 (0.1107)
(500,500)	**0.4882 (0.0282)**	0.4698 (0.0322)	0.4103 (0.0308)	0.4397 (0.0317)	0.432 (0.0343)
High correlation ($\mathbf{\Sigma_1 = \Sigma_2 = 0.3 \cdot I + 0.7 \cdot J}$)					
(50,50)	**0.4392 (0.0904)**	0.3598 (0.1004)	0.2652 (0.106)	0.249 (0.0925)	0.2646 (0.0892)
(500,500)	**0.4653 (0.0286)**	0.4457 (0.0315)	0.2966 (0.0367)	0.2954 (0.0378)	0.2963 (0.0362)
Different Correlation ($\mathbf{\Sigma_1 = 0.3 \cdot I + 0.7 \cdot J, \Sigma_2 = 0.7 \cdot I + 0.3 \cdot J}$)					
(50,50)	**0.392 (0.0933)**	0.3758 (0.0946)	0.3212 (0.0978)	0.3288 (0.0892)	0.3194 (0.1007)
(500,500)	0.4262 (0.031)	**0.4814 (0.0314)**	0.3564 (0.0323)	0.3593 (0.0306)	0.3598 (0.0307)
Negative Correlation ($\rho = -0.1$)					
(50,50)	**0.5074 (0.0808)**	0.4578 (0.0857)	0.4358 (0.0826)	0.4708 (0.0802)	0.461 (0.0841)
(500,500)	**0.5562 (0.0239)**	0.5306 (0.0245)	0.4658 (0.027)	0.5147 (0.0279)	0.5053 (0.0323)

Table 3. Normal distributions. Same means. Four biomarkers.

Size (n_1,n_2)	LR	XG	MM	MMM	MMIQR
Low correlation ($\mathbf{\Sigma_1 = \Sigma_2 = 0.7 \cdot I + 0.3 \cdot J}$)					
(50,50)	**0.4878 (0.09)**	0.429 (0.0942)	0.4794 (0.0954)	0.482 (0.097)	0.4818 (0.0971)
(500,500)	**0.5254 (0.0285)**	0.5125 (0.0284)	0.5073 (0.0278)	0.5183 (0.032)	0.5182 (0.0295)
Different Correlation ($\mathbf{\Sigma_1 = 0.3 \cdot I + 0.7 \cdot J, \Sigma_2 = 0.7 \cdot I + 0.3 \cdot J}$)					
(50,50)	0.4598 (0.0958)	0.4628 (0.1004)	0.5366 (0.0904)	**0.5438 (0.0813)**	0.5398 (0.0869)
(500,500)	0.4783 (0.0259)	0.5279 (0.0284)	0.5609 (0.0285)	0.5627 (0.0271)	**0.5632 (0.0271)**
Different Correlation ($\mathbf{\Sigma_1 = 0.5 \cdot I + 0.5 \cdot J, \Sigma_2 = I}$)					
(50,50)	0.5452 (0.0779)	0.5186 (0.0831)	0.5758 (0.0776)	0.587 (0.0825)	**0.59 (0.0771)**
(500,500)	0.5748 (0.028)	0.5875 (0.0289)	0.5975 (0.0279)	**0.6097 (0.0247)**	0.6088 (0.0246)

Tables 4 and 5 present the results obtained considering the above scenarios for $p = 10$ biomarkers. The code can be found at Supplementary Material: Chapter 1, Sections A.6.–A.10. Table 4 shows the Youden indices achieved for biomarkers with different means. The conclusions derived are similar to those in Table 2, with our more hardened approaches being significantly superior to the min–max approach, especially when biomarkers are independent or negatively correlated, where they achieve the best performance. Logistic regression generally outperforms all other algorithms in all other scenarios.

The results reported in Table 5 show similar behaviour to Table 3. In general, the summary statistics-based methods and, in particular, our approaches outperform the others. The XGBoost algorithm outperforms logistic regression, generally, in the different correlation scenarios.

3.1.2. Asymmetric Distributions

This section presents results derived from simulated scenarios of non-normal distributions. Tables 6–8 show the results obtained from simulated data for $p = 4$ biomarkers. The code can be found in Supplementary Material: Chapter 1, Sections B.1.–B.6. Specifically, Tables 6 and 7 consider scenarios under log-normal distribution and Table 8 considering different marginal distributions (χ^2, normal, gamma and exponential). Tables 9 and 10 display the results obtained from simulated data for $p = 10$ biomarkers following a log-normal distribution. The code can be found in Supplementary Material: Chapter 1, Sections B.7.–B.11.

Table 4. Normal distributions. Different means. Ten biomarkers.

Size (n_1,n_2)	LR	XG	MM	MMM	MMIQR
Independents					
(50,50)	**0.8698 (0.055)**	0.86 (0.061)	0.7704 (0.0642)	0.8716 (0.0547)	0.8616 (0.0647)
(500,500)	**0.9448 (0.0093)**	0.9288 (0.0142)	0.7962 (0.0187)	0.8996 (0.0146)	0.8993 (0.0175)
High correlation ($\boldsymbol{\Sigma_1 = \Sigma_2 = 0.3 \cdot I + 0.7 \cdot J}$)					
(50,50)	**0.8324 (0.0696)**	0.7992 (0.0726)	0.6672 (0.0854)	0.6644 (0.0868)	0.662 (0.0861)
(500,500)	**0.9191 (0.0149)**	0.8935 (0.0168)	0.6893 (0.0245)	0.6899 (0.0253)	0.6888 (0.0271)
Different Correlation ($\boldsymbol{\Sigma_1 = 0.3 \cdot I + 0.7 \cdot J, \Sigma_2 = 0.7 \cdot I + 0.3 \cdot J}$)					
(50,50)	**0.789 (0.071)**	0.7658 (0.0706)	0.4924 (0.0932)	0.4948 (0.0961)	0.4982 (0.0977)
(500,500)	0.8585 (0.0174)	**0.8626 (0.0179)**	0.5287 (0.0274)	0.5471 (0.0255)	0.5464 (0.0248)
Negative Correlation ($\rho = -0.1$)					
(50,50)	0.9232 (0.0552)	0.8884 (0.075)	0.865 (0.0599)	0.9468 (0.0482)	**0.9576 (0.0418)**
(500,500)	0.9936 (0.0042)	0.9764 (0.0145)	0.8953 (0.0166)	0.996 (0.0035)	**0.9962 (0.0034)**

Table 5. Normal distributions. Same means. Ten biomarkers..

Size (n_1,n_2)	LR	XG	MM	MMM	MMIQR
Low correlation ($\boldsymbol{\Sigma_1 = \Sigma_2 = 0.7 \cdot I + 0.3 \cdot J}$)					
(50,50)	0.5216 (0.09)	0.508 (0.0833)	0.5202 (0.0812)	**0.5328 (0.0885)**	0.531 (0.0841)
(500,500)	**0.5803 (0.0264)**	0.5595 (0.0255)	0.5501 (0.0281)	0.5742 (0.0302)	0.5751 (0.0285)
Different Correlation ($\boldsymbol{\Sigma_1 = 0.3 \cdot I + 0.7 \cdot J, \Sigma_2 = 0.7 \cdot I + 0.3 \cdot J}$)					
(50,50)	0.4206 (0.1047)	0.4662 (0.106)	**0.6784 (0.0758)**	0.6774 (0.0821)	0.6766 (0.08)
(500,500)	0.5044 (0.0283)	0.6324 (0.0279)	0.692 (0.0226)	0.6929 (0.0233)	**0.6946 (0.0242)**
Different Correlation ($\boldsymbol{\Sigma_1 = 0.5 \cdot I + 0.5 \cdot J, \Sigma_2 = I}$)					
(50,50)	0.621 (0.0816)	0.6146 (0.0819)	0.6966 (0.0751)	0.7172 (0.0706)	**0.718 (0.0715)**
(500,500)	0.6849 (0.0237)	0.714 (0.0257)	0.712 (0.0203)	0.7373 (0.025)	**0.7378 (0.0253)**

Table 6 reports results for the scenario of biomarkers with different means. It shows tnat logistic regression outperforms the others in high-correlation scenarios. The XGBoost algorithm dominates the others in all other scenarios for larger sample sizes and in all scenarios of different correlations. Our min–max-IQR approach outperforms the others in small sample sizes of negative correlations and particularly the min–max approach in the independent biomarker scenarios.

The results reported in Table 7 show a similar behaviour to Table 3 (normal distributions) but with a worse logistic regression performance, such that our approaches generally perform the best in all scenarios.

Table 6. Log-normal distributions. Different means. Four biomarkers.

Size (n_1,n_2)	LR	XG	MM	MMM	MMIQR
Independents					
(50,50)	**0.4112 (0.0936)**	0.385 (0.1016)	0.3658 (0.106)	0.393 (0.1031)	0.3852 (0.1077)
(500,500)	0.4562 (0.03)	**0.4686 (0.0313)**	0.4096 (0.0324)	0.4376 (0.0315)	0.435 (0.0328)
High correlation ($\mathbf{\Sigma_1 = \Sigma_2 = 0.3 \cdot I + 0.7 \cdot J}$)					
(50,50)	**0.4108 (0.1009)**	0.354 (0.0998)	0.2596 (0.1025)	0.2468 (0.1004)	0.246 (0.102)
(500,500)	**0.4502 (0.0282)**	0.446 (0.0318)	0.2954 (0.0365)	0.2935 (0.037)	0.2935 (0.0337)
Different Correlation ($\mathbf{\Sigma_1 = 0.3 \cdot I + 0.7 \cdot J, \Sigma_2 = 0.7 \cdot I + 0.3 \cdot J}$)					
(50,50)	0.3492 (0.0955)	**0.3814 (0.0879)**	0.3158 (0.0942)	0.3198 (0.0955)	0.3166 (0.102)
(500,500)	0.394 (0.0337)	**0.4815 (0.0328)**	0.3558 (0.0325)	0.3571 (0.0317)	0.3567 (0.0334)
Negative Correlation ($\rho = -0.1$)					
(50,50)	0.4538 (0.0963)	0.4466 (0.0861)	0.4268 (0.0834)	0.4794 (0.0795)	**0.4802 (0.0853)**
(500,500)	0.4916 (0.0264)	**0.5304 (0.0252)**	0.4627 (0.0248)	0.5051 (0.0274)	0.5044 (0.0278)

Table 7. Log-normal distributions. Same means. Four biomarkers.

Size (n_1,n_2)	LR	XG	MM	MMM	MMIQR
Low correlation ($\mathbf{\Sigma_1 = \Sigma_2 = 0.7 \cdot I + 0.3 \cdot J}$)					
(50,50)	0.4592 (0.0955)	0.4312 (0.0943)	0.4704 (0.0933)	**0.483 (0.0879)**	0.4816 (0.095)
(500,500)	0.5049 (0.0301)	0.5121 (0.029)	0.5051 (0.027)	**0.5161 (0.0282)**	0.5151 (0.0282)
Different Correlation ($\mathbf{\Sigma_1 = 0.3 \cdot I + 0.7 \cdot J, \Sigma_2 = 0.7 \cdot I + 0.3 \cdot J}$)					
(50,50)	0.4066 (0.0926)	0.4524 (0.1019)	**0.5444 (0.0911)**	0.5358 (0.0905)	0.54 (0.0884)
(500,500)	0.4166 (0.0279)	0.5292 (0.0278)	0.5606 (0.0266)	**0.5607 (0.0276)**	0.56 (0.0279)
Different Correlation ($\mathbf{\Sigma_1 = 0.5 \cdot I + 0.5 \cdot J, \Sigma_2 = I}$)					
(50,50)	0.4556 (0.0853)	0.522 (0.0755)	0.5784 (0.0813)	**0.5832 (0.0865)**	0.5798 (0.0897)
(500,500)	0.4755 (0.0282)	0.5881 (0.0288)	0.5982 (0.0276)	0.6074 (0.025)	**0.6079 (0.0258)**

The results in Table 8 indicate that, in scenarios of different marginal distributions, the XGBoost algorithm dominates the rest significantly. In these scenarios, the summary statistics-based methods are the worst performers.

Table 8. Different marginal distributions.Four biomarkers.

Size (n_1,n_2)	LR	XG	MM	MMM	MMIQR
(50,50)	0.6572 (0.108)	**0.6838 (0.0947)**	0.3716 (0.1194)	0.3442 (0.1234)	0.3616 (0.1251)
(500,500)	0.7065 (0.0363)	**0.7692 (0.0220)**	0.435 (0.0453)	0.4357 (0.0442)	0.4358 (0.0429)

Table 9 shows that logistic regression outperforms the rest in high-correlation scenarios but is closely followed by the XGBoost algorithm. The XGBoost algorithm dominates over the others in independent scenarios of larger sample sizes and scenarios with different correlations between groups. Our approaches achieve the best performance in independent biomarker scenarios and smaller sample sizes and in scenarios with negative correlations.

Table 9. Log-normal distributions. Different means. Ten biomarkers.

Size (n_1,n_2)	LR	XG	MM	MMM	MMIQR
Independents					
(50,50)	0.8344 (0.0556)	0.8594 (0.065)	0.7728 (0.0632)	**0.8772 (0.0545)**	0.8646 (0.0591)
(500,500)	0.901 (0.0137)	**0.9293 (0.0128)**	0.7966 (0.018)	0.895 (0.0154)	0.8929 (0.0178)
High correlation ($\mathbf{\Sigma_1 = \Sigma_2 = 0.3 \cdot I + 0.7 \cdot J}$)					
(50,50)	**0.817 (0.0706)**	0.8034 (0.0745)	0.6584 (0.0762)	0.6572 (0.0828)	0.657 (0.0805)
(500,500)	**0.8948 (0.0164)**	0.8916 (0.0173)	0.6825 (0.0224)	0.6812 (0.0244)	0.6814 (0.0248)
Different Correlation ($\mathbf{\Sigma_1 = 0.3 \cdot I + 0.7 \cdot J}$, $\mathbf{\Sigma_2 = 0.7 \cdot I + 0.3 \cdot J}$)					
(50,50)	0.7436 (0.0795)	**0.7692 (0.077)**	0.4994 (0.0932)	0.505 (0.0961)	0.5026 (0.097)
(500,500)	0.808 (0.0217)	**0.8626 (0.018)**	0.5218 (0.0279)	0.5452 (0.0257)	0.5445 (0.025)
Negative Correlation ($\rho = -0.1$)					
(50,50)	0.8946 (0.0586)	0.886 (0.0755)	0.8704 (0.05)	0.952 (0.0449)	**0.953 (0.0457)**
(500,500)	0.9703 (0.01)	0.9756 (0.0143)	0.8938 (0.0149)	0.9947 (0.0044)	**0.9964 (0.0036)**

The results reported in Table 10 show a general dominance of our approaches over the others, with logistic regression performing significantly worse in scenarios of different correlation than in other scenarios.

Table 10. Log-normal distributions. Same means.Ten biomarkers.

Size (n_1,n_2)	LR	XG	MM	MMM	MMIQR
Low correlation ($\mathbf{\Sigma_1 = \Sigma_2 = 0.7 \cdot I + 0.3 \cdot J}$)					
(50,50)	0.5072 (0.0917)	0.5066 (0.0859)	0.5144 (0.0869)	0.5306 (0.0931)	**0.5344 (0.0886)**
(500,500)	0.5656 (0.0253)	0.5592 (0.0292)	0.5459 (0.0277)	0.5739 (0.0296)	**0.5748 (0.0288)**
Different Correlation ($\mathbf{\Sigma_1 = 0.3 \cdot I + 0.7 \cdot J}$, $\mathbf{\Sigma_2 = 0.7 \cdot I + 0.3 \cdot J}$)					
(50,50)	0.35 (0.1058)	0.4666 (0.1037)	**0.6732 (0.0852)**	0.661 (0.091)	0.664 (0.0905)
(500,500)	0.4253 (0.0277)	0.6321 (0.0275)	0.6796 (0.0239)	0.6802 (0.0254)	**0.6811 (0.0251)**
Different Correlation ($\mathbf{\Sigma_1 = 0.5 \cdot I + 0.5 \cdot J}$, $\mathbf{\Sigma_2 = I}$)					
(50,50)	0.4944 (0.0835)	0.613 (0.0756)	0.69 (0.0789)	0.7166 (0.0762)	**0.7186 (0.0782)**
(500,500)	0.5506 (0.0265)	0.715 (0.0239)	0.7122 (0.0225)	**0.738 (0.0231)**	0.7376 (0.0222)

3.1.3. Summary

To provide a better understanding of our proposed approaches' performance compared to other algorithms, this section provides a summary of the previously displayed results in Figures 1 and 2. Figure 1 shows the results from simulated scenarios of normal distributions, while Figure 2 displays the results from non-normal distributions. The value on the y-axis represents the value after subtracting the average Youden index achieved by our proposed best approach (MMM or MMMIQR) among the other algorithms: logistic regression, XGBoost and min–max approach. The blue value corresponds to the difference with logistic regression (denoted by MMM-LR), red with XGBoost algorithm (MMM-XG), and black with min–max approach (MMM-MM). Thus, negative values on the graph represent scenarios where algorithms outperform our approaches and positive values in other scenarios. The further away from zero, the more significant the difference.

Regarding scenarios with normal distributions (Figure 1), machine learning algorithms (logistic regression and XGBoost algorithm) outperform our approaches, particularly in scenarios with biomarkers with different predictive capacity, mainly in scenarios with high correlations and different correlations (scenarios 2 and 3). However, our approaches

outperform the others in scenarios with biomarkers with similar predictive ability (scenarios 5–7) and in scenarios of biomarkers with negative correlations, mainly for scenarios with a higher number of biomarkers. Note that these differences are more pronounced in scenarios with a higher number of biomarkers.

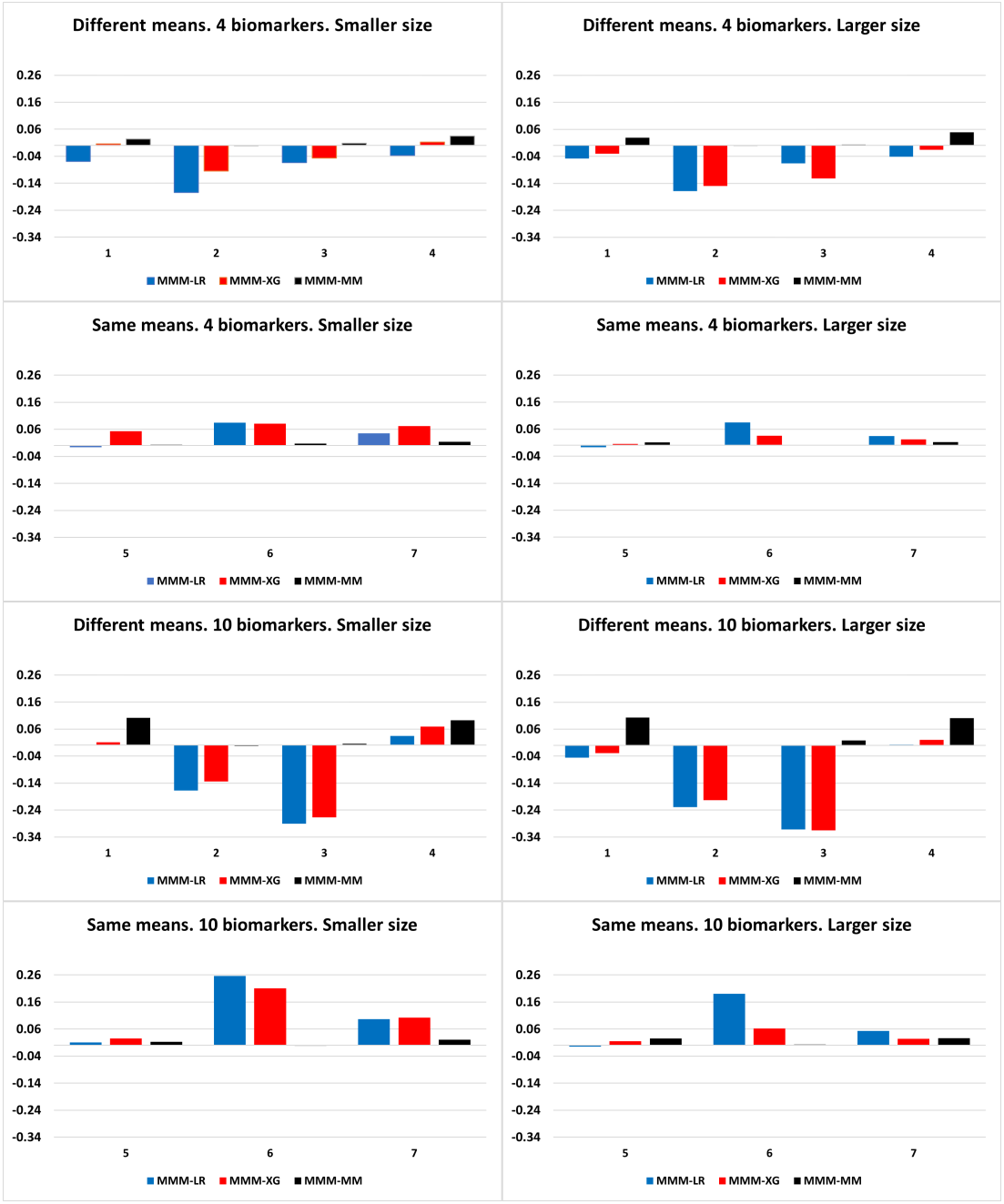


Figure 1. Normal distributions. Difference in the average Youden index achieved by our approach (MMM/MMIQR) and the other algorithms (MMM-LR, MMM-XG, MMM-MM). **1**: Independents. **2**: High correlations. **3**: Different correlations ($\Sigma_1 = 0.3 \cdot I + 0.7 \cdot J, \Sigma_2 = 0.7 \cdot I + 0.3 \cdot J$). **4**: Negative correlations. **5**: Low correlation. **6**: Different correlations ($\Sigma_1 = 0.3 \cdot I + 0.7 \cdot J, \Sigma_2 = 0.7 \cdot I + 0.3 \cdot J$). **7**: Different correlations ($\Sigma_1 = 0.5 \cdot I + 0.5 \cdot J, \Sigma_2 = I$).

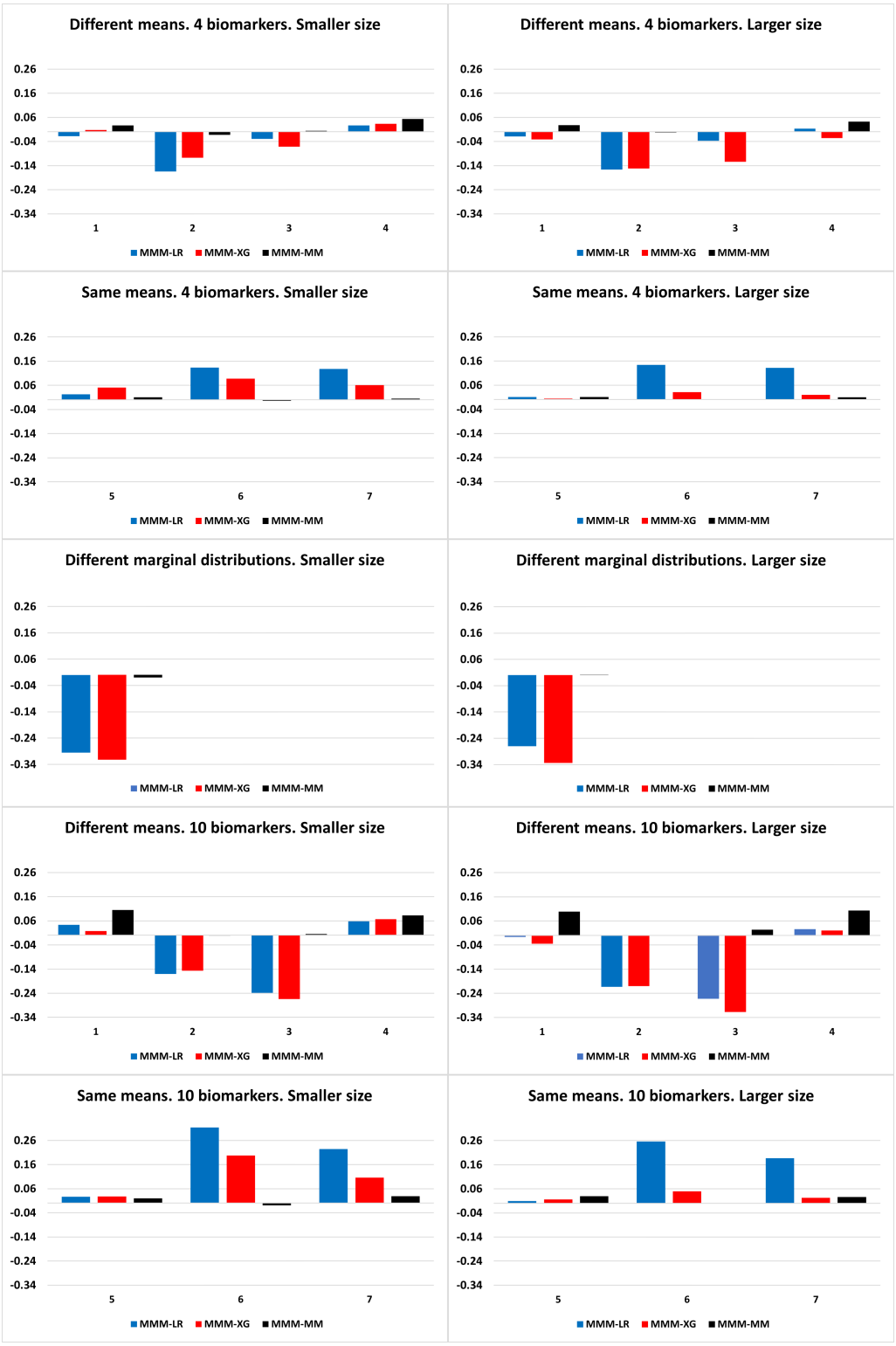


Figure 2. Non-normal distributions. Difference in the average Youden index achieved by our approach (MMM/MMIQR) and the other algorithms (MMM-LR, MMM-XG, MMM-MM). **1**: Log-normal. Independents. **2**: Log-normal. High correlations. **3**: Log-normal. Different correlations ($\Sigma_1 = 0.3 \cdot I + 0.7 \cdot J, \Sigma_2 = 0.7 \cdot I + 0.3 \cdot J$). **4**: Log-normal. Negative correlations. **5**: Log-normal. Low correlation. **6**: Log-normal. Different correlations ($\Sigma_1 = 0.3 \cdot I + 0.7 \cdot J, \Sigma_2 = 0.7 \cdot I + 0.3 \cdot J$). **7**: Log-normal. Different correlations ($\Sigma_1 = 0.5 \cdot I + 0.5 \cdot J, \Sigma_2 = I$).

As for the scenarios of non-normal distributions (Figure 2), the behaviour in scenarios of log-normal distributions is similar to those with normal distributions (Figure 1 above), but with larger differences overall in favour of our algorithm with respect to logistic regression in scenarios with biomarkers of the same means. In scenarios where biomarkers follow different marginal distributions, the XGBoost algorithm significantly achieves the best performance. XGBoost also outperforms the others in scenarios of biomarkers with different means and different variance-covariance matrices between groups.

3.2. Real Datasets

3.2.1. Duchenne Muscular Dystrophy

Figure 3 shows the distribution of each biomarker, and Table 11 displays the correlation matrix between them in each group (67 carriers and 127 non-carriers) for the Duchenne muscular dystrophy dataset, where r_{CK-H} denotes the correlation between the pair of biomarkers CK and H. The code can be found in Supplementary Material: Chapter 2, Section A.1.

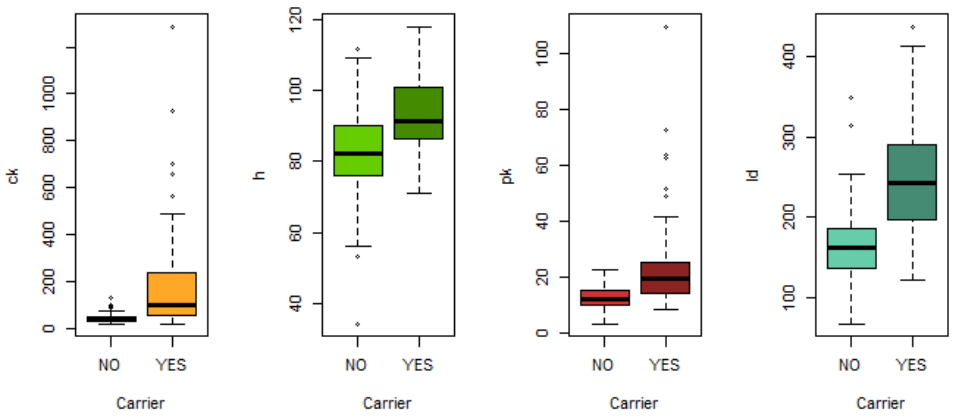


Figure 3. Marginal distributions of biomarkers. DMD dataset.

Table 11. Correlation between biomarkers. DMD dataset.

	r_{CK-H}	r_{CK-PK}	r_{CK-LD}	r_{H-PK}	r_{H-LD}	r_{PK-LD}
Non-Carrier	−0.33	0.1	0.2	0.08	0.18	0.22
Carrier	−0.14	0.7	0.49	−0.12	−0.1	0.48

Because the range of biomarker values differs from the others, the values of each biomarker were normalised before applying summary statistics-based methods, thus ensuring the correct use of these methods. The biomarkers CK and H show a negative correlation, which is stronger in the non-carrier group. Conversely, the other biomarker pairs generally show positive correlations, with a stronger correlation observed in the carrier group.

The estimates of the Youden index of each biomarker (CK, H, PK, LD) produced in a univariate way on the whole dataset were 0.612, 0.417, 0.508, and 0.578. Table 12 presents the average value of the maximum Youden indices achieved in each fold for each of the analysed methods, as well as their respective values of sensitivity and specificity. The code can be found in Supplementary Material: Chapter 2, Section A.2.

Logistic regression achieved the best performance, although our approaches are not far behind, outperforming the XGBoost algorithm and notably the min–max approach.

3.2.2. Maternal Health Risk

This section presents the performance results of the approaches to the problem of predicting high or medium versus low risk of maternal mortality (High–Medium vs. Low Risk) and the problem of predicting high risk versus low or medium risk of maternal mortality (High vs. Medium–Low Risk) for the Maternal Health Risk dataset.

Table 12. Ten-fold cross-validation. DMD dataset.

Algorithm	Youden	Sensitivity	Specificity
LR	**0.8008**	0.8476	0.9532
XG	0.7047	0.8142	0.8905
MM	0.6258	0.7952	0.8306
MMM	0.7793	0.8595	0.9198
MMIQR	0.7772	0.8761	0.9011

3.2.3. High–Medium vs. Low Risk

Figure 4 displays the distribution of each variable, and Table 13 shows the correlation matrix between them in each group (196 low risk and 181 medium-high risk). The variables Age, SystolicBP, DiastolicBP, BS, BodyTemp and HeartRate are denoted by V1, V2, V3, V4, V5, and V6, respectively. The code can be found in Supplementary Material: Chapter 2, Section B.1.

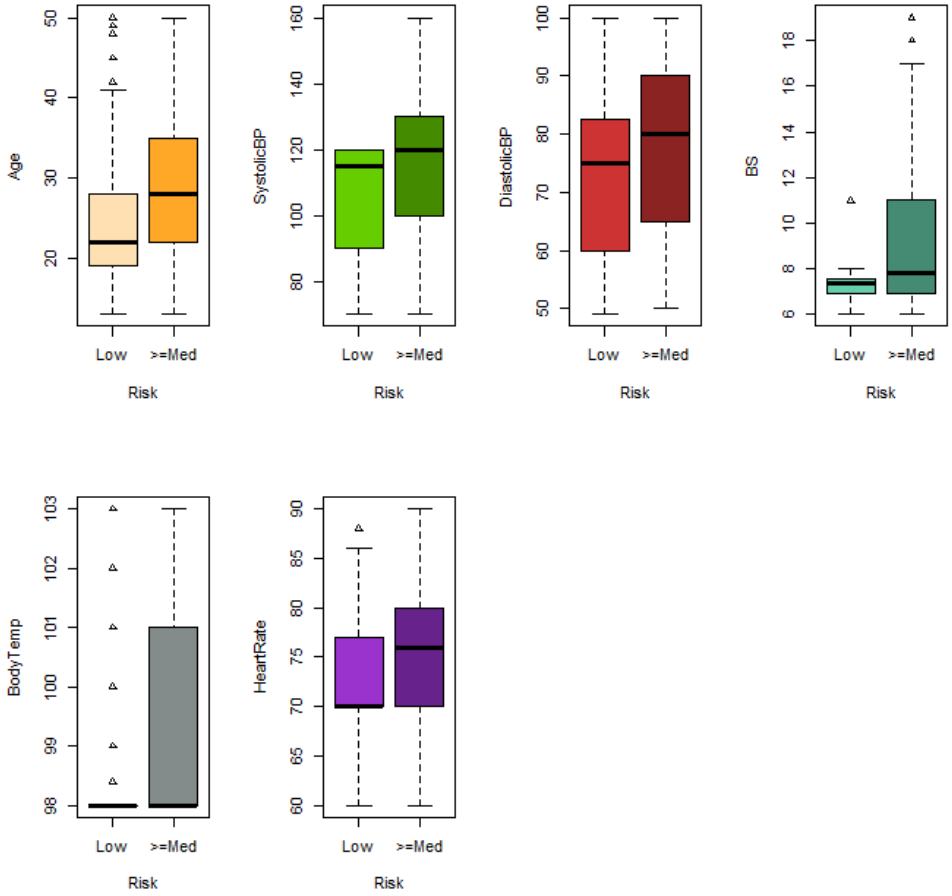


Figure 4. Marginal distributions of biomarkers. Maternal Health dataset. High–Medium vs. Low Risk.

Table 13. Correlation between biomarkers. Maternal Health dataset. High–Medium vs. Low Risk.

	r_{V1-V2}	r_{V1-V3}	r_{V1-V4}	r_{V1-V5}	r_{V1-V6}	r_{V2-V3}	r_{V2-V4}	r_{V2-V5}	r_{V2-V6}	r_{V3-V4}	r_{V3-V5}	r_{V3-V6}	r_{V4-V5}	r_{V4-V5}	r_{V5-V6}
Low Risk	0.39	0.4	0.17	−0.13	−0.17	0.8	0.07	−0.08	−0.15	0.11	−0.1	−0.09	0.03	−0.02	0.11
High–Medium Risk	0.46	0.39	0.53	−0.38	0.08	0.75	0.27	−0.41	−0.07	0.27	−0.36	−0.14	−0.17	0.19	0.12

Positive and negative correlations are shown between the variables, generally with greater strength in the higher-risk group (High–Medium Risk). The estimates of the Youden index of each biomarker (Age, SystolicBP, DiastolicBP, BS, BodyTemp and HeartRate)

produced in a univariate way on the whole dataset were 0.305, 0.304, 0.153, 0.377, 0.249, 0.217, respectively. The predictive capacity of the biomarkers in this dataset was lower than the previously presented DMD dataset, as can also be seen in Figure 4.

Table 14 presents the performance achieved by each of the analysed methods after the application of 10-fold cross-validation. The values of each biomarker were normalised before applying the summary statistics-based methods. The code can be found in Supplementary Material: chapter 2, section B.2.

Table 14. Ten-fold cross-validation. Maternal Health dataset. High–Medium vs.Low Risk.

Algorithm	Youden	Sensitivity	Specificity
LR	0.5366	0.6611	0.8755
XG	**0.586**	0.74	0.846
MM	0.4563	0.7055	0.7508
MMM	0.4739	0.718	0.7559
MMIQR	0.4739	0.718	0.7559

The XGBoost algorithm and logistic regression achieved better performance than the summary-statistics-based methods, especially XGBoost which performs the best.

3.2.4. High vs. Medium–Low Risk

Figure 5 shows the distribution of each variable and Table 15 the correlation matrix between them in each group (282 low–medium risk and 95 high risk). Age, SystolicBP, DiastolicBP, BS, BodyTemp and HeartRate are denoted by V1, V2, V3, V4, V5 and V6, respectively. The code can be found in Supplementary Material: Chapter 2, Section C.1.

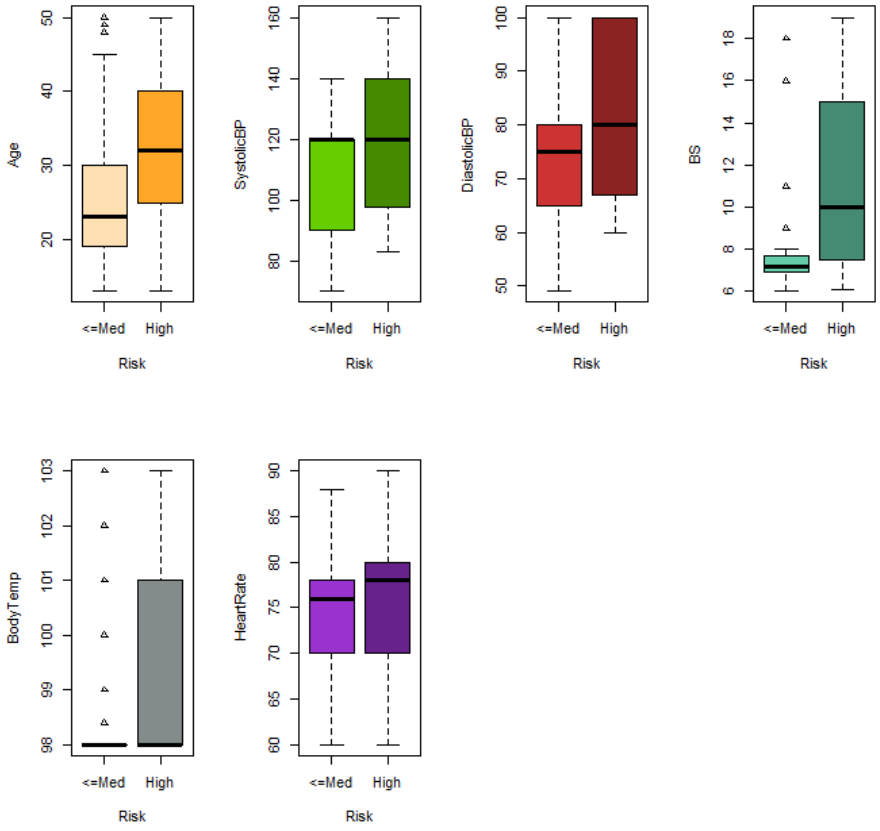


Figure 5. Marginal distributions of biomarkers. Maternal Health dataset. High vs. Medium–Low Risk.

Table 15. Correlation between biomarkers. Maternal Health dataset. High vs. Medium–Low Risk.

	r_{V1-V2}	r_{V1-V3}	r_{V1-V4}	r_{V1-V5}	r_{V1-V6}	r_{V2-V3}	r_{V2-V4}	r_{V2-V5}	r_{V2-V6}	r_{V3-V4}	r_{V3-V5}	r_{V3-V6}	r_{V4-V5}	r_{V4-V5}	r_{V5-V6}
Medium–Low Risk	0.44	0.41	0.33	−0.14	−0.13	0.74	0.18	−0.08	−0.16	0.17	−0.1	−0.16	0.05	0.03	0.22
High Risk	0.44	0.4	0.55	−0.57	0.06	0.84	0.23	−0.57	−0.04	0.18	−0.54	−0.09	−0.4	0.14	−0.01

As in the previous example, positive and negative correlations between pairs of variables are shown. The Youden index estimates for each biomarker (Age, SystolicBP, DiastolicBP, BS, BodyTemp and HeartRate) univariate over the whole dataset were 0.347, 0.386, 0.268, 0.564, 0.22 and 0.272, respectively, which are higher than in the previous example.

Table 16 displays the performance achieved for each of the analysed methods. The values of each biomarker were normalised before applying the summary statistics-based methods. The code can be found in Supplementary Material: Chapter 2, Section C.2.

Table 16. Ten-fold cross-validation. Maternal Health dataset. High vs. Medium–Low Risk.

Algorithm	Youden	Sensitivity	Specificity
LR	0.6225	0.8036	0.849
XG	**0.7159**	0.8238	0.8921
MM	0.6149	0.8932	0.7217
MMM	0.5831	0.8851	0.6981
MMIQR	0.5831	0.8851	0.6981

The XGBoost algorithm significantly outperformed the other algorithms. It was followed by logistic regression, but the summary-statistics-based algorithm was not far behind.

4. Discussion and Conclusions

In binary classification problems in healthcare, the choice of thresholds to dichotomise the model output into groups of patients is crucial and can aid decision-making in clinical practice. In the absence of consensus on the benefits of optimising the classification of one group or another, the Youden index is a standard criterion that provides good performance for the model.

Models combining biomarkers for binary classification have received sufficient attention in the literature. The parametric approach has the limitation of meeting the assumption of normality; by contrast, other authors propose non-parametric approaches without assumptions of biomarker distributions but with the limitation of being computationally intractable when the number of biomarkers increases.

Liu et al. [19] proposed the min–max approach, which is a non-parametric and computationally tractable approach regardless of the number of biomarkers, based on the linear combination of minimum and maximum values of biomarkers. The idea behind this proposal is that the maximum and minimum values chosen among the biomarkers can allow the best discrimination between sick and healthy patients. However, this approach may not be sufficient in terms of discrimination when the number of biomarkers grows, because it is not enough to capture all the discrimination abilities of the set of predictor variables. To improve the min–max algorithm, we proposed the min–max–median/IQR approach under a Youden index maximisation that incorporates a new summary statistic with reasonably good performance. This approach uses a stepwise algorithm that we proposed in [25], which is based on the work of Pepe et al. [14,15].

The use of machine learning algorithms, such as XGBoost, has become increasingly popular in recent years due to their ease of implementation and good results. However, the choice of the optimal approach depends on the problem and the data to be processed. It is therefore essential to make a thorough comparison before providing some guidelines for the selection of the optimal algorithm.

The aim of this paper is to present a comprehensive comparison of our min–max–median/IQR approaches with the min–max approach and machine learning algorithms such as logistic regression and XGBoost, in order to optimise the Youden index. For this purpose, the algorithms were compared on 59 different simulated data scenarios with symmetric and non-symmetric distributions, as well as on two real-world datasets.

The results of the simulated scenarios showed that the machine learning approaches outperformed our approaches, in particular in scenarios with biomarkers with different predictive abilities and in biomarker scenarios with different marginal distributions. However, our approaches outperformed them in scenarios with biomarkers with normal and log-normal distributions with the same predictive ability and different correlations between groups.

Regarding the real datasets, XGBoost outperformed the other algorithms in predicting maternal health risk, while logistic regression achieved the best performance in predicting Duchenne dystrophy, with our proposed approaches closely following. The data show that the problem of predicting Duchenne dystrophy is simpler than that of predicting maternal death risk. In the former, linear combination approaches outperform XGBoost. However, XGBoost outperforms the others on the more complex problem. This may be due to its ability to capture non-linear relationships.

In summary, regardless of the symmetry assumption, non-parametric approaches are always a good alternative for modelling data, but their performance is not guaranteed. Therefore, the modelling process requires a combination of techniques and the optimization of hyperparameters , as we have demonstrated with extensive simulations and application on real data. This work provides the scientific community with a comparison of the performance of our approaches (min–max–median/IQR) and machine learning algorithms, that can be applied and explored in different binary classification problems, such as cancer diagnosis. We proposed a non-parametric approach, addressing the limitations of previous linear biomarker combinations that assume multivariate normality. It also addresses the limitation of the computational burden of certain approaches in the literature, being always approachable regardless of the number of initial biomarkers. This is achieved thanks to the formulation of our algorithm that linearly combines the minimum, maximum, and median or interquartile range biomarkers, thereby converting the n-biomarker combination problem into a three-biomarker combination problem. Although there are several techniques that reduce the dimensionality of the problem for subsequent classification algorithm application, our proposed approach provides a different perspective in this regard. The way our approach is formulated allows the three biomarkers considered (minimum, maximum and median or interquartile range) to correspond to different original biomarkers for each patient. This offers the possibility of capturing biomarker heterogeneity in the data. Subsequently, our approach applies a stepwise algorithm, which we published in [25] , and which demonstrated acceptable performance in the comparison study. Therefore, our approach proposes a novel formulation in the state of the art that addresses certain limitations in the literature. Furthermore, a comparison of our approach with other approaches, such as the XGBoost algorithm, provides performance results with other approaches that also help to capture biomarker heterogeneity, albeit from a different perspective. XGBoost is a decision tree ensemble algorithm that builds multiple trees and combines their predictions to produce the final output. For the construction of each tree, a random subset of biomarkers/variables is selected and used to partition the data. In this way, each tree is constructed with information from different biomarkers, thus helping to avoid overfitting.

Our approaches have been shown to be superior to other algorithms, including machine learning algorithms, in scenarios with biomarkers having the same predictive capacity and different correlations between groups. These results are not surprising, as there is a variety of health problems in which the combination of the minimum and maximum of biomarkers provides the best classification. In prostate cancer, the worst diagnosis corresponds to a higher value in PSA and a lower value in prostate volume. PSA density

is defined by the division of PSA and prostate volume, and it shows a better predictive ability than PSA. Thus, we can choose as a biomarker for prostate cancer the PSA density or the combination provided by PSA and prostate volume. Moreover, as with PSA, there is a variety of competing biomarkers such as PCA3, SelectMdx, and 4Kscore, in which high values correspond to a greater probability of cancer. The min–max derived approaches gives the opportunity to choose from them the one which takes the highest value. Similar to prostate volume, the free PSA takes lower values with a worse diagnosis of prostate cancer; therefore, choosing the minimum and maximum marker for a group of candidates with similar performance can contribute to the best discrimination ability. In addition, the third parameter, median or interquartile range, informs about the performance of the set of biomarkers. The cost-effectiveness of a set of biomarkers to diagnose a unique disease can be controversial, but molecular or metabolomic markers are associated with a variety of cancers, and their analysis has been increasing in recent years. The stratification of cancer or its prognosis will be derived from biomarkers built from information derived from different perspectives.

Although our work includes an exhaustive comparison study in various real and simulated data scenarios, yielding interesting results, the conclusions must be considered within the framework of our study. All conclusions derived from our study are limited to the scenarios and algorithms explored. One of the limitations of the study is the variety of machine learning algorithms considered. While the XGBoost algorithm and logistic regression have been widely used in recent years and have proven efficient, a comparative study that includes additional machine-learning techniques would provide more consistent conclusions. In future work, we propose exploring other machine learning algorithms, deep learning and ensemble models, to compare their performance with our approaches, particularly in scenarios where they are optimal.

Another limitation of the study is the variety of real datasets used, where in no case did our approach achieve the best performance. As future work, we propose evaluating the performance of our approach on real datasets that meet the conditions of the optimal simulation scenarios. One example could be the dataset used in [19], where the authors demonstrated that the min–max combination of three growth hormones (IGFBP3, IGF1, and GHBP) was superior to the other linear combinations for identifying autism. The aim of this evaluation would be to determine whether our approaches outperform min–max and the other algorithms studied.

In addition to scenarios combining multiple biomarkers with the same predictive capability, our approach could also be applied in scenarios where repeated measurements of a single biomarker are recorded, converting the temporal information into three summary measurements. Readers are encouraged to evaluate our approaches in such problems, for example, the detection of events or neurodegenerative diseases from gait information retrieved from wearable sensor measurements. Another line of future work could involve adapting the models and the study to other objective metrics, such as the weighted Youden index.

In conclusion, our study presents a comprehensive comparison of various approaches, presenting our proposed approach (min–max–Median/IQR) as an alternative to machine learning models such as logistic regression and XGBoost, in certain scenarios where it has demonstrated superior performance. We believe that the results of this research will provide valuable insights for the development and application of classification algorithms in the field of medicine, such as cancer diagnosis.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/sym15030756/s1.

Author Contributions: Conceptualisation, R.A.-G. and L.M.E.; methodology, R.A.-G., L.M.E., G.S. and R.d.-H.A.; software, R.A.-G. and L.M.E. and R.d.-H.-A.; validation, R.A.-G. and L.M.E.; formal analysis, R.A.-G. and L.M.E.; investigation, R.A.-G., L.M.E. and G.S.; resources, R.A.-G. and L.M.E.; data curation, R.A.-G. and L.M.E.; writing—original draft preparation, R.A.-G. and L.M.E.;

writing—review and editing, R.A.-G., L.M.E., G.S. and R.d.-H.-A; visualisation, R.A.-G. and L.M.E.; supervision, R.A.-G., L.M.E., G.S. and R.d.-H.-A.; funding acquisition, R.d.-H.-A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was made possible through the funding of project MIA.2021.M02.0007 of the NextGenerationEU program and the support of the Integration and Development of Big Data and Electrical Systems (IODIDE) group of the Aragon Government program. L.M. Esteban and G. Sanz were supported by Gobierno de Aragón (E46-20R) and Ministerio de Ciencia e Innovación (PID2020-116873GB-I00).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Duchenne muscular dystrophy dataset can be found at https://hbiostat.org/data/, accessed on 30 January 2023. The Maternal Health Risk dataset can be found at http://archive.ics.uci.edu/ml, accessed on 30 January 2023.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ROC	Receiver operating characteristic
AUC	Area under the ROC curve
XGBoost	Extreme Gradient Boosting
SVM	Support Vector Machine
XAI	Explainable artificial intelligence
MM	Min–max approach
MMM	Min–max–median approach
MMIQR	Min–max-IQR approach
IQR	Interquartile range
LR	Logistic regression
XGB	XGBoost algorithm
DMD	Duchenne Muscular Dystrophy
CK	Serum creatine kinase
H	haemopexin
PK	Pyruvate kinase
LD	Lactate dehydrogenase
SDG	Sustainable Development Goals
SystolicBP	Upper value of blood pressure
DiastolicBP	Lower value of blood pressure
BS	Blood glucose levels
BodyTemp	Body temperature
HeartRate	Normal resting heart rate

References

1. Pinsky, P F.; Zhu, C.S. Building multi-marker algorithms for disease prediction—The role of correlations among markers. *Biomark. Insights* **2011**, 6, BMI-S7513.
2. Bansal, A.; Pepe, M.S. When does combining markers improve classification performance and what are implications for practice? *Stat. Med.* **2013**, *32*, 1877–1892.
3. Esteban, L.M.; Sanz, G.; Borque, A. Linear combination of biomarkers to improve diagnostic accuracy in prostate cancer. *Monogr. MatemáTicas GarcíA Gald.* **2013**, *38*, 75–84.
4. Kang, L.; Xiong, C.; Crane, P.; Tian, L. Linear combinations of biomarkers to improve diagnostic accuracy with three ordinal diagnostic categories. *Stat. Med.* **2013**, *32*, 631–643.
5. Yan, L.; Tian, L.; Liu, S. Combining large number of weak biomarkers based on AUC. *Stat. Med.* **2015**, *34*, 3811–3830.
6. Amini, M.; Kazemnejad, A.; Zayeri, F.; Amirian, A.; Kariman, N. Application of adjusted-receiver operating characteristic curve analysis in combination of biomarkers for early detection of gestational diabetes mellitus. *Koomesh* **2019**, *21*, 751–758.
7. Ahmadian, R.; Ercan, I.; Sigirli, D.; Yildiz, A. Combining binary and continuous biomarkers by maximizing the area under the receiver operating characteristic curve. *Commun. Stat. Simul. Comput.* **2022**, *51*, 4396–4409.

8. Gargallo-Puyuelo, C.J.; Aznar-Gimeno, R.; Carrera-Lasfuentes, P.; Lanas, A.; Ferrandez, A.; Quintero, E.; Carrillo, M.; Alonso-Abreu, I.; Esteban L.M.; Rodrigálvarez-Chamarro, M.V.; et al. Predictive Value of Genetic Risk Scores in the Development of Colorectal Adenomas. *Dig. Dis. Sci.* **2022**, *67*, 4049–4058.
9. Pastor-Navarro, B.; Rubio-Briones, J.; Borque-Fernando, A.; Esteban, L.M.; Dominguez-Escrig, J.L.; Lopez-Guerrero, J.A. Active Surveillance in Prostate Cancer: Role of Available Biomarkers in Daily Practice. *Int. J. Mol. Sci.* **2021**, *22*, 6266. https://doi.org/10.3390/ijms22126266
10. Faraggi, D.; Reiser, B. Estimation of the area under the ROC curve. *Stat. Med.* **2002**, *21*, 3093–3106.
11. Youden, W.J. Index for rating diagnostic tests. *Cancer J.* **1950**, *3*, 32–35.
12. Su, J.Q.; Liu, J.S. Linear combinations of multiple diagnostic markers. *J. Am. Stat. Assoc.* **1993**, *88*, 1350–1355.
13. Capitanio, U.; Perrotte, P.; Zini, L.; Suardi, N.; Antebi, E.; Cloutier, V.; Jeldres, C.; Shariat, S.F.; Duclos, A.; Arjane, P.; et al. Population-based analysis of normal Total PSA and percentage of free/Total PSA values: Results from screening cohort. *Urology* **2009**, *73*, 1323–1327. https://doi.org/10.1016/j.urology.2008.10.026.
14. Pepe, M.S.; Thompson, M.L. Combining diagnostic test results to increase accuracy. *Biostatistics* **2000**, *1*, 123–140.
15. Pepe, M.S.; Cai, T.; Longton, G. Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics* **2006**, *62*, 221–229.
16. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36.
17. Esteban, L.M.; Sanz, G.; Borque, A. A step-by-step algorithm for combining diagnostic tests. *J. Appl. Stat.* **2011**, *38*, 899–911.
18. Kang, L.; Liu, A.; Tian, L. Linear combination methods to improve diagnostic/prognostic accuracy on future observations. *Stat. Methods Med. Res.* **2016**, *25*, 1359–1380.
19. Liu, C.; Liu, A.; Halabi, S. A min–max combination of biomarkers to improve diagnostic accuracy. *Stat. Med.* **2011**, *30*, 2005–2014.
20. Mi, G.; Li, W.; Nguyen, T.S. Characterize and Dichotomize a Continuous Biomarker. In *Statistical Methods in Biomarker and Early Clinical Development*; Springer: Cham, Switzerland, 2019; pp. 23–38.
21. Perkins, N.J., Schisterman, E.F. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *Am. J. Epidemiol.* **2006**, *163*, 670–675.
22. Martínez-Camblor, P., Pardo-Fernández, J.C. The Youden Index in the Generalized Receiver Operating Characteristic Curve Context. *Int. J. Biostat.* **2019**, *15*, 20180060.
23. Lopes Vendrami, C.; McCarthy, R.J.; Chatterjee, A.; Casalino, D.; Schaeffer, E.M.; Catalona, W.J.; Miller, F.H. The Utility of Prostate Specific Antigen Density, Prostate Health Index, and Prostate Health Index Density in Predicting Positive Prostate Biopsy Outcome is Dependent on the Prostate Biopsy Methods. *Urology* **2019**, *129*, 153–159. https://doi.org/10.1016/j.urology.2019.03.018.
24. Yin, J.; Tian, L. Optimal linear combinations of multiple diagnostic biomarkers based on Youden index. *Stat. Med.* **2014**, *33*, 1426–1440.
25. Aznar-Gimeno, R.; Esteban, L.M.; del-Hoyo-Alonso, R.; Borque-Fernando, Á.; Sanz, G. A Stepwise Algorithm for Linearly Combining Biomarkers under Youden Index Maximization. *Mathematics* **2022**, *10*, 1221.
26. Walker, S.H.; Duncan, D.B. Estimation of the probability of an event as a function of several independent variables. *Biometrika* **1967**, *54*, 167–179.
27. Aznar-Gimeno, R.; Esteban, L. M.; Sanz, G.; del-Hoyo-Alonso, R.; Savirón-Cornudella, R.; Antolini L. Incorporating a New Summary Statistic into the Min–Max Approach: A Min–Max–Median, Min–Max–IQR Combination of Biomarkers for Maximising the Youden Index. *Mathematics* **2021**, *9*, 2497.
28. Sarker, I.H. Machine learning: Algorithms, real-world applications and research directions. *SN Comput. Sci.* **2021**, *2*, 1–21.
29. Fatima, M.; Pasha, M. Survey of machine learning algorithms for disease diagnostic. *J. Intell. Learn. Syst. Appl.* **2017**, *9*, 1.
30. Nilashi, M.; bin Ibrahim, O.; Ahmadi, H.; Shahmoradi, L. An analytical method for diseases prediction using machine learning techniques. *Comput. Chem. Eng.* **2017**, *106*, 212–223.
31. Sidey-Gibbons, J.A.; Sidey-Gibbons, C.J. Machine learning in medicine: A practical introduction. *BMC Med. Res. Methodol.* **2019**, *19*, 1–18.
32. Aznar-Gimeno, R.; Esteban, L.M.; Labata-Lezaun, G.; del-Hoyo-Alonso, R.; Abadia-Gallego, D.; Paño-Pardo, J.R.; Esquillor-Rodrigo, M.J.; Lanas, A.; Serrano, M.T. A clinical decision web to predict ICU admission or death for patients hospitalised with COVID-19 using machine learning algorithms. *Int. J. Environ. Res. Public Health* **2021**, *18*, 8677.
33. Pappada, S.M. Machine learning in medicine: It has arrived, let's embrace it. *J. Card. Surg.* **2021**, *36*, 4121–4124.
34. Navarro, C.L.A.; Damen, J.A.; van Smeden, M.; Takada, T.; Nijman, S.W.; Dhiman, P.; Ma, J.; Collins, G.S.; Bajpai, R.; Riley, R.D.; et al. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. *J. Clin. Epidemiol.* **2022**, 154, 8-22.
35. Agrawal, S.; Jain, S.K. Medical text and image processing: Applications, issues and challenges. *Mach. Learn. Health Care Perspect. Mach. Learn. Healthc.* **2020**, 13, 237–262.
36. Shehab, M.; Abualigah, L.; Shambour, Q.; Abu-Hashem, M.A.; Shambour, M.K.Y.; Alsalibi, A.I.; Gandomi, A.H. Machine learning in medical applications: A review of state-of-the-art methods. *Comput. Biol. Med.* **2022**, *145*, 105458.
37. Amethiya, Y.; Pipariya, P.; Patel, S.; Shah, M. Comparative analysis of breast cancer detection using machine learning and biosensors. *Intell. Med.* **2022**, *2*, 69–81.

38. Riyaz, L.; Butt, M.A.; Zaman, M.; Ayob, O. Heart disease prediction using machine learning techniques: A quantitative review. In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021*; Springer: Singapore, 2022; Volume 3, pp. 81–94.
39. Huang, S.; Yang, J.; Shen, N.; Xu, Q.; Zhao, Q. Artificial intelligence in lung cancer diagnosis and prognosis: Current application and future perspective. In *Seminars in Cancer Biology*; Academic Press: Cambridge, MA, USA, 2023
40. Nematollahi, H.; Moslehi, M.; Aminolroayaei, F.; Maleki, M.; Shahbazi-Gahrouei, D. Diagnostic Performance Evaluation of Multiparametric Magnetic Resonance Imaging in the Detection of Prostate Cancer with Supervised Machine Learning Methods. *Diagnostics* **2023**, *13*, 806.
41. Aznar-Gimeno, R.; Labata-Lezaun, G.; Adell-Lamora, A.; Abadia-Gallego, D.; del-Hoyo-Alonso, R.; Gonzalez-Muñoz, C. Deep learning for walking behaviour detection in elderly people using smart footwear. *Entropy* **2021**, *23*, 777.
42. Poirion, O.B.; Jing, Z.; Chaudhary, K.; Huang, S.; Garmire, L. DeepProg: An ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. *Genome Med.* **2021**, *13*, 112.
43. Grapov, D.; Fahrmann, J.; Wanichthanarak, K.; Khoomrung, S. Rise of deep learning for genomic, proteomic, and metabolomic data integration in precision medicine. *OMICS* **2018**, *22*, 630–636.
44. Alakwaa, F.M.; Chaudhary, K.; Garmire, L.X. Deep learning accurately predicts estrogen receptor status in breast cancer metabolomics data. *J. Proteome Res.* **2018**, *17*, 337–347.
45. Mahesh, T.R.; Vinoth Kumar, V.; Muthukumaran, V.; Shashikala, H.K.; Swapna, B.; Guluwadi, S. Performance Analysis of XGBoost Ensemble Methods for Survivability with the Classification of Breast Cancer. *J. Sens.* **2022**. https://doi.org/10.1155/2022/4649510
46. Botlagunta, M.; Botlagunta, M.D.; Myneni, M.B.; Lakshmi, D.; Nayyar, A.; Gullapalli, J.S.; Shah, M.A. Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms. *Sci. Rep.* **2023**, *13*, 485.
47. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
48. Rustam, Z.; Darmawan, N.A.; Hartini, S.; Aurelia, J.E. (2022). Support Vector Machines and Naïve Bayes Classifier for Classifying a Prostate Cancer. In *Advanced Intelligent Systems for Sustainable Development (AI2SD'2020)*; Springer International Publishing: Cham, Switzerland, 2022; Volume 1, pp. 854–860.
49. Huo, X.; Finkelstein, J. Prostate Cancer Prediction Using Classification Algorithms 2022. Available online: (accessed on 14 March 2023).
50. Sabbagh, A.; Washington, S.L., III; Tilki, D.; Hong, J.C.; Feng, J.; Valdes, G.; Chen, M-H.; Wu, J.; Huland H.; Graefen, M.; et al. Development and External Validation of a Machine Learning Model for Prediction of Lymph Node Metastasis in Patients with Prostate Cancer. *Eur. Urol. Oncol.* **2023**. https://doi.org/10.1016/j.euo.2023.02.006
51. Khan, A.; Tariq, I.; Khan, H.; Khan, S.U.; He, N.; Zhiyang, L.; Raza, F. Lung Cancer Nodules Detection via an Adaptive Boosting Algorithm Based on Self-Normalized Multiview Convolutional Neural Network. *J. Oncol.* **2022**, *2022*, 5682451.
52. Saheb-Honar, M.; Dehaki, M.G.; Kazemi-Galougahi, M.H.; Soleiman-Meigooni, S. A Comparison of Three Research Methods: Logistic Regression, Decision Tree, and Random Forest to Reveal Association of Type 2 Diabetes with Risk Factors and Classify Subjects in a Military Population. *JAMM* **2022**, *10*.
53. Budholiya, K.; Shrivastava, S.K.; Sharma, V. An optimized XGBoost based diagnostic system for effective prediction of heart disease. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 4514–4523.
54. Mathema, V.B.; Sen, P.; Lamichhane, S.; Orešič, M.; Khoomrung, S. Deep learning facilitates multi-data type analysis and predictive biomarker discovery in cancer precision medicine. *Comput. Struct. Biotechnol. J.* **2023**, *21*, 1372–1382. https://doi.org/10.1016/j.csbj.2023.01.043.
55. Tran, K.A.; Kondrashova, O.; Bradley, A.; Williams, E.D.; Pearson J.V.; Waddell, N. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Med*. **2021**, *13*, 152. https://doi.org/10.1186/s13073-021-00968-x
56. Kleppe, A.; Skrede, O.J.; De Raedt, S.; Liestøl, K.; Kerr, D.J.; Danielsen, H.E. Designing deep learning studies in cancer diagnostics. *Nat. Rev. Cancer* **2021**, *21*, 199–211. https://doi.org/10.1038/s41568-020-00327-9.
57. Holzinger, A.; Langs, G.; Denk, H., Zatloukal, K.; Müller, H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2019**, *9*, e1312.
58. Bentéjac, C.; Csörgő, A.; Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* **2021**, *54*, 1937–1967.
59. Zhang, Y.; Wang, Y.; Xu, J.; Zhu, B.; Chen, X.; Ding, X.; Li, Y. Comparison of prediction models for acute kidney injury among patients with hepatobiliary malignancies based on XGBoost and lasso-logistic algorithms. *Int. J. Gen. Med.* **2021**, *14*, 1325.
60. Feng, Y.N.; Xu, Z.H.; Liu, J.T.; Sun, X.L.; Wang, D.Q.; Yu, Y. Intelligent prediction of RBC demand in trauma patients using decision tree methods. *Mil. Med. Res.* **2021**, *8*, 1–12.
61. Xiang, L.; Wang, H.; Fan, S.; Zhang, W.; Lu, H.; Dong, B.; Liu, S.; Chen, Y.; Wang, Y.; Zhao, L.; Fu, L. Machine Learning for Early Warning of Septic Shock in Children With Hematological Malignancies Accompanied by Fever or Neutropenia: A Single Center Retrospective Study. *Front. Oncol.* **2021**, *11*, 678743.
62. Larsson, A.; Berg, J.; Gellerfors, M.; Gerdin Wärnberg, M. The advanced machine learner XGBoost did not reduce prehospital trauma mistriage compared with logistic regression: A simulation study. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 1–9.
63. Yan, Z.; Wang, J.; Dong, Q.; Zhu, L.; Lin, W.; Jiang, X. XGBoost algorithm and logistic regression to predict the postoperative 5-year outcome in patients with glioma. *Ann. Transl. Med.* **2022**, *10*, 860.

64. Moore, A; Bell, M. XGBoost, A Novel Explainable AI Technique, in the Prediction of Myocardial Infarction: A UK Biobank Cohort Study. *Clin. Med. Insights Cardiol.* **2022**, *16*, 11795468221133611.
65. Wang, R.; Wang, L.; Zhang, J.; He, M.; Xu, J. XGBoost Machine Learning Algorism Performed Better Than Regression Models in Predicting Mortality of Moderate-to-Severe Traumatic Brain Injury. *World Neurosurg.* **2022**, *163*, e167–e622.
66. de Hond, A.A.; Kant, I.M.; Honkoop, P.J.; Smith, A.D.; Steyerberg, E.W.; Sont, J.K. Machine learning did not beat logistic regression in time series prediction for severe asthma exacerbations. *Sci. Rep.* **2022**, *12*, 1–8.
67. Volovici, V.; Syn, N.L.; Ercole, A.; Zhao, J.J.; Liu, N. Steps to avoid overuse and misuse of machine learning in clinical research. *Nat. Med.* **2022**, *28*, 1996–1999.
68. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020–2021. Available online: http://www.r-project.org/index.html (accessed on 9 January 2023).
69. SLModels: Stepwise Linear Models for Binary Classification Problems under Youden Index Optimisation. R Package Version 0.1.2. Available online: https://cran.r-project.org/web/packages/SLModels/index.html (accessed on 9 January 2023).
70. Percy, M.E.; Andrews, D.F.; Thompson, M.W. Duchenne muscular dystrophy carrier detection using logistic discrimination: Serum creatine kinase, hemopexin, pyruvate kinase, and lactate dehydrogenase in combination. *Am. J. Med. Genet.* **1982**, *13*, 27–38.
71. Dua, D.; Graff, C. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science 2019. Available online: http://archive.ics.uci.edu/ml (accessed on 30 January 2023).

Capítulo 7

Estimación en problemas de clasificación en entornos clínicos reales

Los enfoques lineales y los métodos de Machine Learning son herramientas valiosas para abordar problemas clínicos reales, como el diagnóstico de enfermedades, la predicción de eventos y la identificación de factores de riesgo. Estos métodos permiten estimar modelos clasificatorios que pueden detectar los patrones complejos de los datos, lo que puede ayudar a los profesionales de la salud a tomar decisiones más informadas así como contribuir a la mejora del sistema de salud. Seleccionar el modelo adecuado es esencial y depende del contexto y los datos que se estén analizando. Para tomar esta decisión, se debe tener en cuenta una métrica de evaluación específica a optimizar. Además de la capacidad discriminatoria del modelo, la aplicabilidad práctica de estos modelos radica también en su utilidad clínica. En este contexto, el análisis del punto de corte óptimo adquiere una relevancia significativa al trasladar al clínico el resultado del modelo en información interpretable con una dicotomización en clases.

En los capítulos anteriores, se han presentado trabajos donde se han comparado métodos de clasificación utilizando tanto simulaciones como aplicaciones con datos reales en medicina. Continuando en esta dirección, en las siguientes secciones presentamos el resto de trabajos incluidos en este compendio de tesis [5, 34, 97, 3], los cuales se centran en la estimación y aplicación de modelos en problemas de clasificación en entornos clínicos reales. Estos resultados tienen un impacto positivo en la práctica clínica al proporcionar herramientas para la toma de decisiones médicas.

7.1. Predicción del riesgo en pacientes con COVID-19

El trabajo [5] presentado en esta sección se enmarca en el contexto de la pandemia de COVID-19, donde la evaluación del riesgo en pacientes positivos no era evidente en la práctica clínica. Nuestro objetivo fue construir un modelo de predicción del riesgo de ingreso en la UCI o muerte en pacientes hospitalizados con COVID-19, para integrarlo en una aplicación web fácil de usar para ayudar al clínico en la toma de decisiones rápidas, de especial valor en tiempos de pandemia.

Nuestro estudio abordó ciertas limitaciones de trabajos relacionados del estado del arte. Exploramos técnicas de aprendizaje automático como los algoritmos ensemble (Random Forest y XGBoost) y algoritmos de redes neuronales. La evaluación de los modelos se realizó tanto interna como externamente y su calibración fue también evaluada. Para la selección del mejor modelo se llevó a cabo una optimización de los hiperparámetros del modelo con el fin de seleccionar la arquitectura óptima. Asimismo, se abordaron aspectos relacionados con la utilidad clínica.

Con el fin de facilitar su aplicación clínica, se seleccionaron las 20 variables con mayor capacidad predictiva de entre las 150 originales, basadas en la información del paciente en el momento del ingreso hospitalario. Esta elección fue robusta puesto que no resultó en una pérdida de capacidad predictiva, como demostramos con test de comparación de curvas ROC. El mejor modelo, en términos de AUC, se logró con el algoritmo XGBoost (AUC=0.82). Debido a la falta de interpretabilidad intrínseca de los modelos de XGBoost, se aplicaron técnicas de explicabilidad para ofrecer una interpretación de las decisiones del modelo.

Además de una correcta validación del modelo, se llevó a cabo un análisis de los puntos de corte óptimos según diferentes criterios, incluyendo el accuracy, sensibilidad, especificidad, valor predictivo positivo, valor predictivo negativo y el índice de Youden. La selección de un punto de corte óptimo favorece la implementación de los modelos clasificatorios en la práctica clínica, ya que permite al profesional elegir el umbral más adecuado.

El resultado final fue una aplicación web sencilla que integraba los modelos estimados. A partir de la información del paciente (20 variables explicativas), la herramienta devuelve la probabilidad de que el paciente termine en una condición grave (UCI o muerte) y ofrece una representación gráfica que permite una interpretación de los resultados.

A continuación se presenta el artículo, donde se explica el proceso seguido en mayor profundidad así como los resultados obtenidos.

International Journal of
Environmental Research and Public Health

MDPI

Perspective

A Clinical Decision Web to Predict ICU Admission or Death for Patients Hospitalised with COVID-19 Using Machine Learning Algorithms

Rocío Aznar-Gimeno [1,*], Luis M. Esteban [2], Gorka Labata-Lezaun [1], Rafael del-Hoyo-Alonso [1,*], David Abadia-Gallego [1], J. Ramón Paño-Pardo [3,4,5], M. José Esquillor-Rodrigo [6], Ángel Lanas [4,5,7,8] and M. Trinidad Serrano [4,5,7]

1 Department of Big Data and Cognitive Systems, Instituto Tecnológico de Aragón, ITAINNOVA, María de Luna 7-8, 50018 Zaragoza, Spain; glabata@itainnova.es (G.L.-L.); dabadia@itainnova.es (D.A.-G.)
2 Escuela Universitaria Politécnica de La Almunia, Universidad de Zaragoza, Calle Mayor, 5, 50100 La Almunia de Doña Godina, Spain; lmeste@unizar.es
3 Infectious Disease Department, University Clinic Hospital Lozano Blesa, San Juan Bosco 15, 50009 Zaragoza, Spain; joserrapa@gmail.com
4 Department of Medicine, Psychiatry and Dermatology, University of Zaragoza, 50009 Zaragoza, Spain; alanas@unizar.es (Á.L.); tserrano.aullo@gmail.com (M.T.S.)
5 Aragon Health Research Institute (IIS Aragon), 50009 Zaragoza, Spain
6 Internal Medicine Department, University Clinic Hospital Lozano Blesa, San Juan Bosco 15, 50009 Zaragoza, Spain; mjesquillor@gmail.com
7 Digestive Diseases Department, University Clinic Hospital Lozano Blesa, San Juan Bosco 15, 50009 Zaragoza, Spain
8 CIBEREHD, 28029 Madrid, Spain
* Correspondence: raznar@itainnova.es (R.A.-G.); rdelhoyo@itainnova.es (R.d.-H.-A.)

Citation: Aznar-Gimeno, R.; Esteban, L.M.; Labata-Lezaun, G.; del-Hoyo-Alonso, R.; Abadia-Gallego, D.; Paño-Pardo, J.R.; Esquillor-Rodrigo, M.J.; Lanas, Á.; Serrano, M.T. A Clinical Decision Web to Predict ICU Admission or Death for Patients Hospitalised with COVID-19 Using Machine Learning Algorithms. *Int. J. Environ. Res. Public Health* **2021**, *18*, 8677. https://doi.org/10.3390/ijerph18168677

Academic Editor: Jimmy T. Efird

Received: 6 July 2021
Accepted: 11 August 2021
Published: 17 August 2021

Abstract: The purpose of the study was to build a predictive model for estimating the risk of ICU admission or mortality among patients hospitalized with COVID-19 and provide a user-friendly tool to assist clinicians in the decision-making process. The study cohort comprised 3623 patients with confirmed COVID-19 who were hospitalized in the SALUD hospital network of Aragon (Spain), which includes 23 hospitals, between February 2020 and January 2021, a period that includes several pandemic waves. Up to 165 variables were analysed, including demographics, comorbidity, chronic drugs, vital signs, and laboratory data. To build the predictive models, different techniques and machine learning (ML) algorithms were explored: multilayer perceptron, random forest, and extreme gradient boosting (XGBoost). A reduction dimensionality procedure was used to minimize the features to 20, ensuring feasible use of the tool in practice. Our model was validated both internally and externally. We also assessed its calibration and provide an analysis of the optimal cut-off points depending on the metric to be optimized. The best performing algorithm was XGBoost. The final model achieved good discrimination for the external validation set (AUC = 0.821, 95% CI 0.787–0.854) and accurate calibration (slope = 1, intercept = −0.12). A cut-off of 0.4 provides a sensitivity and specificity of 0.71 and 0.78, respectively. In conclusion, we built a risk prediction model from a large amount of data from several pandemic waves, which had good calibration and discrimination ability. We also created a user-friendly web application that can aid rapid decision-making in clinical practice.

Keywords: COVID-19; ICU; mortality; machine learning; predictive model; clinical decision web tool

1. Introduction

1.1. Context

Recent advances in the field of artificial intelligence have demonstrated their success in different fields of interest such as the environment [1], climate change [2], agriculture [3], industry [4] and health [5–7], among others. In particular, the application of modelling and

the development of machine learning algorithms have been emerging in recent times and have been responsible for these multiple applications of interest for data-driven decision support and profit maximisation.

Deep Learning is a subset of machine learning algorithms that has shown great promise especially in the application of data with some temporality [8,9], image or textual data in areas such as image recognition [10–13], natural language processing [14] or speech recognition [15].

These algorithms generally outperform machine learning techniques when the data size is large. This explains the success in these areas where large amounts of information are generated, such as time series data retrieved by sensors in real time or images or textual knowledge generated through the Internet. In addition to this, the success is also partly due to the fact that deep learning algorithms allow automatic feature extraction based on the data, unlike shallow machine learning algorithms whose features must be previously identified by a subject matter expert.

Deep learning algorithms and in particular deep neural networks allow through their layered architecture to extract from more general to more specific features, without the need for domain expertise. Therefore, when we have a large amount of information and/or unstructured data and/or domain knowledge is lacking, deep learning algorithms outperform others as there is no need to worry about feature engineering.

In addition to these advantages, deep learning techniques have the ability to adapt to different domains more easily [4], for example through transfer learning by using pre-trained deep neural networks.

However, as they are more complex techniques, training deep learning algorithms involve a higher computational cost and therefore requires more powerful infrastructure. Although it has been shown that Deep Learning sometimes outperforms classical machine learning techniques, when the data size is not too large, the latter are preferable [16].

In addition to having a lower computational cost and being recommended for smaller datasets, another advantage of shallow machine learning algorithms is that they are easier to interpret and understand compared to Deep Learning where deep networks are trained as a "black box". This interpretability may be the main argument why many sectors use machine learning techniques as opposed to Deep Learning. This may be the case in the field of health and medicine, when the aim is to provide tools to clinicians helping them make critical decisions about a patient based on information about the patient at a specific time, which is the scope of the work we present here.

The application of artificial intelligence techniques in the field of health and medicine has been particularly accentuated in recent months to create problem-solving strategies due to the COVID-19 (Coronavirus disease 2019) pandemic [17].

The COVID-19 pandemic has already affected more than 174 million people, causing nearly 3 million deaths and overloading healthcare systems worldwide [18]. The clinical severity of the disease is highly variable. Most cases are asymptomatic or mild, but 14% of patients have severe disease, and 5% are critical [19]. The case fatality rate in China has been reported to be 2.3%, and in Europe it reaches 4–4.5% [20].

Multiple predictive and prognostic factors associated with increased severity of infection have been described, and models have been built to assist healthcare systems prioritize medical attention [21–23]. Nevertheless, most of these models have significant biases [23]. In addition, the sample sizes have been limited and most models lack external validation or calibration. There have also been several pandemic waves, in which both mortality rate and severity have differed [24], so it would be necessary for models to prove useful in different scenarios over time.

Machine learning has become an advanced tool for diagnosing health problems and predicting the severity and mortality of different diseases. During the COVID-19 pandemic, machine learning tools have been used for a variety of investigations, ranging from image analysis for diagnosis of SARS-CoV-2 pneumonia to predicting future pandemic waves [17]. These algorithms have the advantage of being able to combine a large amount

of information, extracting the most predictive characteristics of the diagnosis associated with coronavirus disease [17,21–23,25–31].

Therefore, during the COVID-19 pandemic, Machine Learning has been crucial in developing tools for a variety of investigations, ranging from image analysis for diagnosing SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) pneumonia, predicting future pandemic waves, detecting multiple predictive and prognostic factors associated with increased severity of infection, to building severity models that help health systems prioritise care [8,21–30].

In the present study, we focus on assessing the prediction of intensive care unit (ICU) admission or mortality risk (severity risk) among patients admitted to the hospital with COVID-19 using machine learning algorithms.

1.2. Related Work and Limitations

Any predictive model must contemplate certain good practices for model building and validation that provide generalizability and easy interpretation. The first thing that must be clearly defined is the target population, as well as the prediction horizon (outcome), so that the model can be validated with good reliability. A lack of or incorrect definition can lead to inappropriate use of the model and biased interpretation of the results. Limited sample sizes are also a known problem when building prediction models, which also implies a high risk of bias and overfitting of the model [32].

Several studies have built predictive models and analysed mortality risk in patients with suspected or confirmed COVID-19, but with an unclear definition of the target population, without specifying the outcome (mortality) period, and at high risk of biases [33–36]. This bias can lead to miscalibrations and overestimation of the discrimination performance, especially when they are not externally validated.

Yan et al. [34] (mortality risk in validated or suspected COVID-19 inpatients) and Shi et al. [36] (death or severe COVID-19 in inpatients with confirmed COVID-19 at admission) validated their models with samples of patients with a different severity than the training set, which can lead to bad calibration. In particular, Yan et al. used temporal validation, selecting only severe cases and Shi et al. validated using less severe cases.

Yue et al. [37] validated their prognostic models (hospital stay of more than 10 days per COVID admission) using 5-fold cross-validation. Although this type of validation is adequate (cross-validation) in order to avoid the overfitting, external validation ensures better generalization of the model, though it is a more difficult type of validity to achieve. Gong et al. [38] (severe COVID-19 infection within minimum 15 days in inpatients with confirmed COVID-19 at admission) and Xie et al. [39] (mortality in hospital in inpatients with confirmed COVID-19 at admission) performed external validation and achieved good performance, though their validation was performed for different centres, possibly due to being early prognostic model studies of the pandemic (with a population from China); the populations studied were patients admitted between January and March 2020 (first pandemic wave). However, subsequent studies [24] have shown that there were differences between the pandemic waves in terms of patient characteristics and severity. In addition to model discrimination, calibration should also be assessed, though this is not always done or not done correctly. Xie et al. [39] (in-hospital mortality of inpatients with confirmed COVID-19 at admission) performed validation and assessed the calibration correctly.

Many recent studies have explored machine learning techniques for the generation of a COVID-19 prognostic model to predict disease severity, demonstrating their great potential [17,25–27] and the ability of these techniques to select the most significant/influential variables [28–30]. Variable selection can be crucial for the generation of a prognostic model in such a medical and emergency setting. It is well known good practice to select the simplest possible model that still achieves an acceptable level of performance [40]. This is even more important in the medical domain, where a limited number of observations [23] but a large number of variables may be available, which may lead to overfitting. Finally, COVID-19 prognostic models of disease severity ultimately aim to provide a user-friendly

tool for clinical practice that can aid rapid decision-making [22,26,28]. This implies that, for its use to be feasible in practice, the model should not include a large number of variables to be introduced.

Yao et al. [27] built a model for detecting COVID-19 disease severity using the support vector machine (SVM) algorithm by finally selecting 28 features (clinical information and blood/urine test data). They used a dataset of 137 patients hospitalized between January and February 2020. Marcos et al. [28] explored several machine learning algorithms (regularized logistic regression, random forest, XGBoost) to identify early patients who will die or require mechanical ventilation during hospitalization. For training and internal validation, they used a cohort of 918 confirmed COVID-19 patients admitted to the Salamanca hospital between March and April 2020 and a cohort of 252 COVID patients from another hospital (admitted to the Barcelona clinical hospital between February and April 2020). To develop a user-friendly and practical calculator, the number of features used by the machine learning model was reduced from 140 to fewer than 10 (demographic variables, comorbidities, chronic medical treatment, clinical characteristics, physical examination parameters, and biochemical parameters). With a similar aim, Patel et al. [30] developed models for predicting the need for intensive care and mechanical ventilation using a cohort of 212 patients. To this end, they considered various machine learning techniques (random forest, MLP, support vector machines, gradient boosting, extra tree classifier, adaboost). Knight et al. [22] developed and validated a pragmatic risk score (4C mortality score) based on eight variables (age, sex, number of comorbidities, respiratory rate, peripheral oxygen saturation, level of consciousness, urea level, and C-reactive protein) with the aim of predicting mortality in patients admitted to the hospital with COVID-19. For this purpose, they used a large patient dataset including patients from 260 hospitals. In particular, model training was performed on a cohort of patients recruited between 6 February and 20 May 2020, and validation was performed on a second cohort of patients recruited between 21 May and 29 June 2020.

1.3. Contributions of Our Work

The aim of our study was to build a model for predicting the risk of ICU or death in hospitalised patients with COVID-19 and integrate it into an easy-to-use web application to aid rapid decision-making in clinical practice, which is very important in these times of pandemic. To this end, machine learning techniques were applied following a good model building and validation exercise.

As described in the previous section, some related studies have some limitations or shortcomings related to different aspects related to the definition of the study population, the sample size, the validation and generalisation of the model, the assessment of the calibration of the model or its practical application with the development of a web application for decision making. Our work addresses all these limitations.

Our model was built with a larger sample size than most of the studies reviewed, covering more temporal information, including several pandemic waves. The model was validated both internally and externally and its calibration was properly assessed. In addition to a correct validation of the model and good model performance values, we developed a user-friendly tool for quick decision making, which is ultimately the real goal in practice. In order to make it feasible to use in practice, the most important input variables were selected.

Specifically, starting from a set of more than 150 clinical variables based on the patient's history and analysis carried out at hospital admission, we built a predictive model using machine learning algorithms with a reasonable dimensionality, including the 20 most predictive variables, which makes it easy to apply from a clinical perspective, that were also validated in a subsequent wave. In addition, we provide a user-friendly tool for practical use.

The following sections present the data retrieval and pre-processing process of the patient information, the statistical techniques and the process of generation and validation of the model carried out. Subsequently, the results are shown and finally the results

obtained are discussed and compared with the state of the art and the main conclusions of our work are drawn.

2. Materials and Methods

2.1. Patient Information Recruitment

This study involved data from patients with SARS-CoV-2 infections confirmed by RT-PCR (Reverse Transcription Polymerase Chain Reaction) who were hospitalized in the SALUD hospital network of Aragon (Spain), which comprises 23 hospitals, between February 2020 and January 2021. The selection criterion was hospitalization within the first 20 days after, and no more than 10 days before, the first positive SARS-CoV-2 PCR test.

Patient information was retrieved through the BIGAN Gestión Clínica platform of the Department of Health of Aragon. This platform accesses the Aragon Health Records Database, which is the primary data source of SALUD, containing demographic and clinical information on patients. The original database included 7498 patients and 165 variables. The outcome studied was ICU admission or mortality within 30 days of hospital admission, which we will refer to as an unfavourable outcome or severity. Each patient's demographics, comorbidities, and medication prescribed in the 6 months prior to hospitalization were considered. Vital signs were recorded upon arrival at the emergency room. Laboratory variables were measured in the first 24 h. This laboratory information was only available for the largest two hospitals of the SALUD network, which accounted for more than 60% of all patients admitted with COVID-19 in the Aragon region during the study period.

To avoid bias due to missing patient information, which could erroneously and unrealistically influence the performance of the model, patients with <65% of the variables filled in were removed from the original database (listwise deleiton). Furthermore, outliers caused by possible human error in filling in the data were analysed and removed. After removing patients with more than 35% of the variables not filled in and outliers and erroneous data, the univariate mean imputation was carried out for the missing data (missing data were imputed as the mean value of the variable). We used imputation techniques with the purpose of a minimum loss in sample size, although we discarded multiple imputation regression methods due to the complexity of the data with a large number of variables.

Our analysis included 3623 patients.

2.2. Statistical Analysis

For model building, the pre-processed dataset with information from February 2020 to November 2020 was considered and split into three separate datasets as is usual in machine learning algorithms: 75% for the training model (fitting the model), 12.5% for model validation (selecting the best model configuration [hyperparameter set] from a set of candidates), and 12.5% for model testing (providing unbiased evaluation metrics that give a generalized value of the performance of the fitted model). In addition, a dataset including information from November 2020 to January 2021 was used for external validation of the final model in a different temporal scenario. We recruited the 626 cases distributed during the period shown in Figure 1 (green color) that verify the study inclusion criteria, these data are completely different to the generation model data and thus provide a validation on a different dataset not used for the building models.

Figure 1 shows the number of hospitalizations reported in the time period considered. Note that the model included information from two different waves of the pandemic, and the external validation included information from yet a different new wave.

For both datasets, we analysed descriptive variables by severity groups (unfavourable/favourable outcome). Medians and interquartile ranges were used for quantitative variables, and absolute and relative frequencies for categorical variables. Variables were compared between severity groups using Mann-Whitney and chi-squared tests as appropriate.

Figure 2 shows a flowchart of the data retrieval and preparation process that summarises the above.

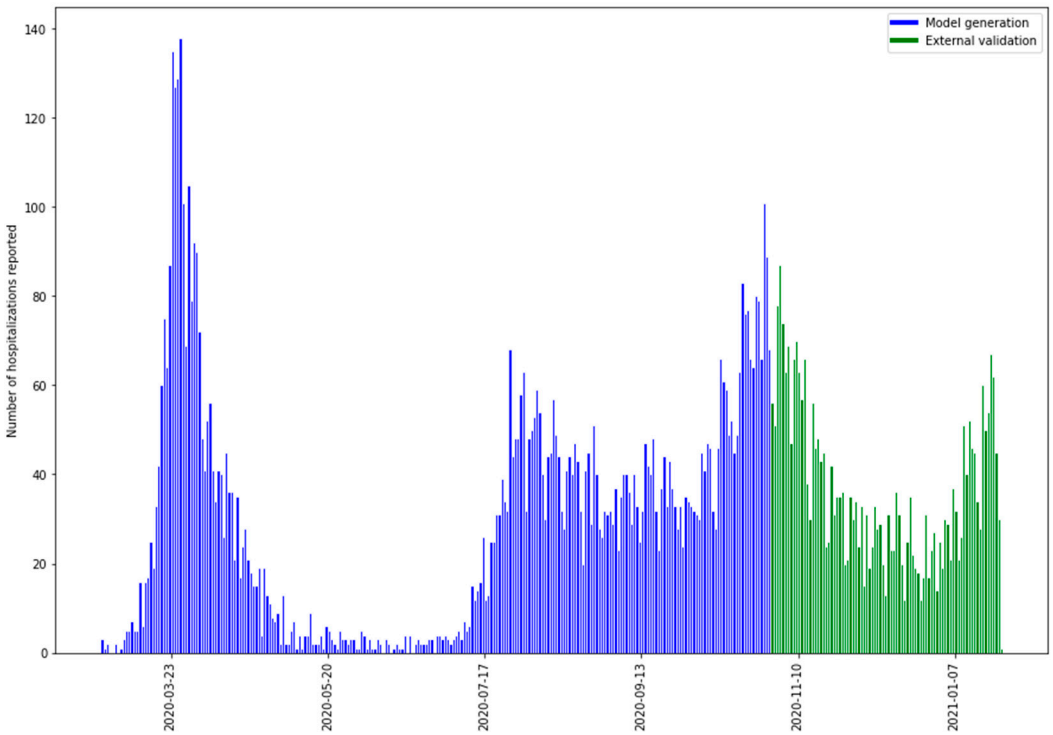


Figure 1. Distribution of the number of hospitalizations reported between February 2020 and January 2021.

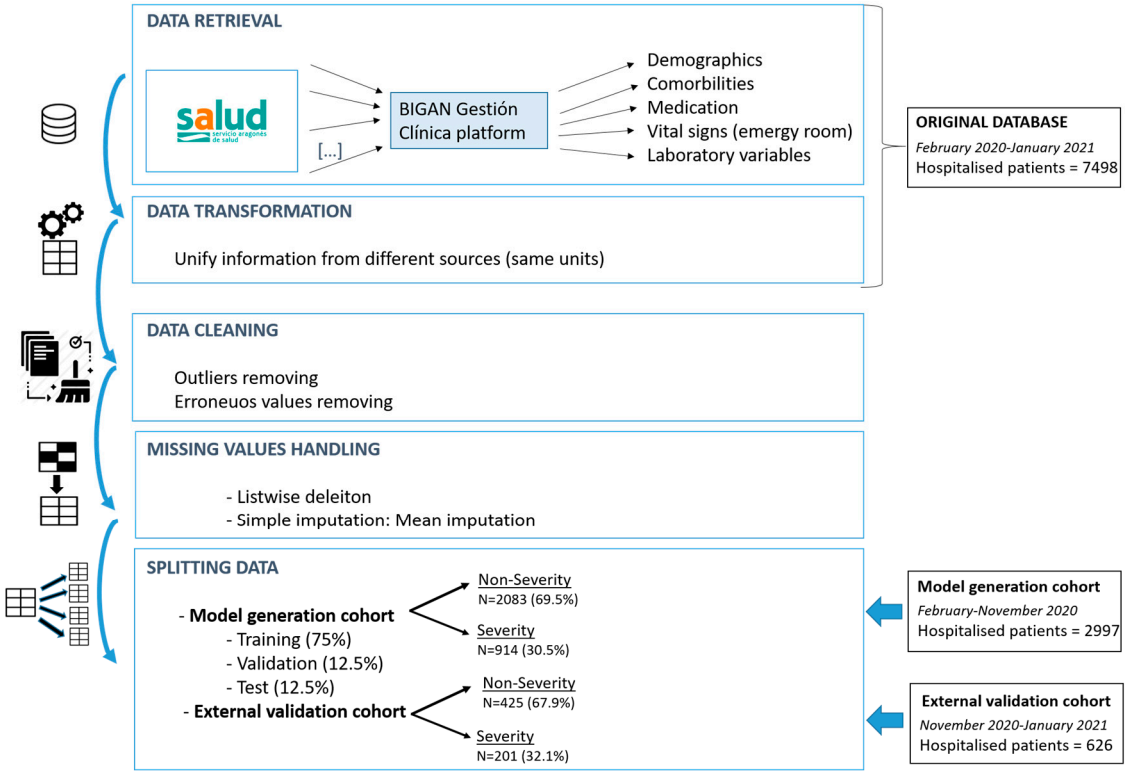


Figure 2. Flowchart of data preparation process.

2.3. Training and Validation

The final aim of the study was to implement a severity diagnostic model in a web tool/platform accessible by clinicians. Thus, by entering the necessary patient information, the probability of severity is provided. In practice, a user-friendly predictive tool such

as this is limited by the amount of patient information (variables) entered; too much information is not feasible in real-world use. Therefore, it was necessary to apply techniques to reduce candidate variables.

Model building was performed in several steps. First, different machine learning algorithms were explored, including artificial neural networks (multilayer perceptron [MLP]), random forest, and gradient boosting trees. The gradient boosting tree algorithms, particularly extreme gradient boosting (XGBoost) [41], achieved the best performance in predicting unfavourable outcomes (ICU admission or death) and were chosen for model development. In addition, ensembles of decision tree methods have the advantage of providing estimates of the importance of variables from a trained predictive model. To report the model, the final importance of each variable is calculated as the average of the importance in each tree, which is calculated by the amount that each attribute split point improves the performance measure weighted by the number of observations for which the node is responsible [42]. This value provides a ranking of feature importance that was used for the selection and reduction of variables.

Second, a stepwise procedure was followed to reduce the number of variables (simpler models) so that the loss in performance was not significant. Models were initially trained using all 165 variables. After the best model with all variables was selected, i.e., the model with the highest discriminatory ability over the validation set, the importance of the variables was calculated and the 50 most important variables selected by means of the XGBoost algorithm. The process was repeated for the dataset using the 50 variables, selecting the 20 most significant variables. The final model was generated by using the dataset with 20 variables.

The discriminative ability of the models was assessed by the area under the receiver operating characteristics (ROC) curve (AUC). The 95% confidence intervals (CIs) were obtained using 2000 stratified bootstrap replicates. The tests for comparing the AUCs proposed by DeLong et al. [43] and Pepe et al. [44] were applied to assess the discriminatory difference between models. In the comparison between models, we considered as the best model the one that corresponds to the largest AUC value. The AUC values can be interpreted as 0.5 = this suggests no discrimination, so we might as well flip a coin. 0.5–0.7 = we consider this poor discrimination, not much better than a coin toss. 0.7–0.8 = acceptable discrimination. 0.8–0.9 = excellent discrimination. >0.9 = Outstanding discrimination [45].

Regarding the model build, the model hyperparameters were optimized over a set of possibilities, choosing the best possible combination (best model) with the selected validation set in order to avoid overfitting. We implemented our model in the Optuna framework to achieve this goal. Optuna [46] is a define-by-run API that allows users to construct the parameter search space dynamically and implements both searching and pruning strategies. In our case, we used the Tree-structured Parzen Estimator (TPE) algorithm [47].

Figure 3 shows a flowchart summarises the above information representing the methodology carried out for the final model generation.

2.4. External Validation

An external dataset with information from November 2020 to January 2021 (external validation set) was used to validate the final model. The calibration (agreement between the probabilities predicted by the model and the real incidence) and discriminatory capacity of the model were analysed, as well as the performance of the model for each cut-off point through the following statistical metrics: accuracy (1), specificity (2), sensitivity (3), Youden (4), positive predictive value (PPV) (5), and negative predictive value (NPV) (6):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (2)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{Youden index} = \text{Sensitivity} + \text{Specificity} - 1 \quad (4)$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (5)$$

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} \quad (6)$$

where TP, TN, FP and FN are the number of true positives, true negatives, false positives and false negatives, respectively.

This clinical utility analysis provides the clinician with decision-making capability, offering a set of different possibilities for the optimal threshold depending on the criterion to be optimized, which may change over time and with circumstances.

The level of significance in the study was established at $p < 0.05$. Analyses were performed using R language programming v 4.0.3 [48] and Python language programming v 3.7.7. [49].

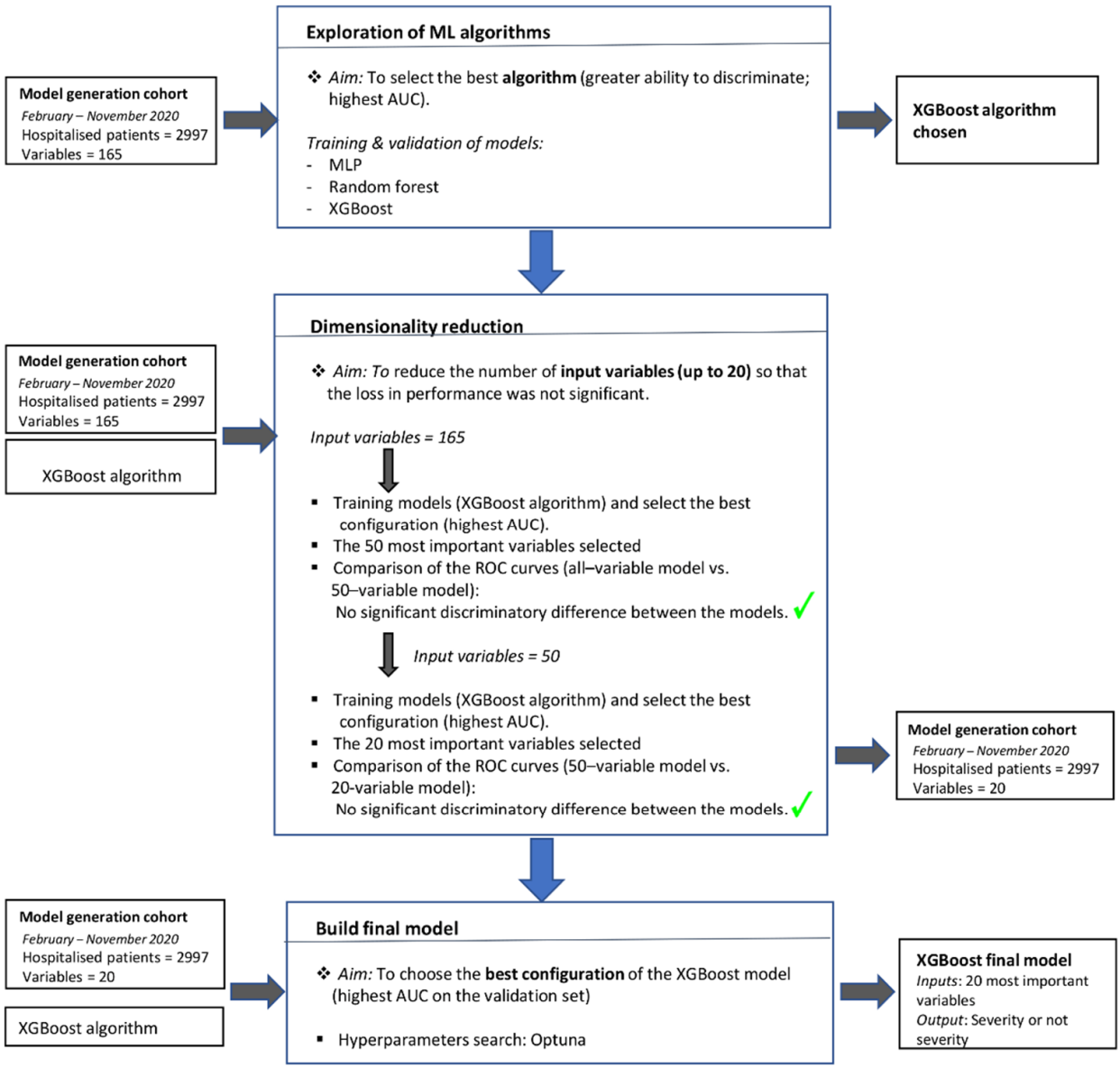


Figure 3. Flowchart of the final model generation process.

3. Results

The descriptive characteristics of the model generation and external validation datasets are provided in Tables 1 and 2, respectively. Significant differences between groups based on the severity were observed for predictive variables in the model generation cohort, and for most of them in the validation cohort.

Table 1. Demographic and clinical information of patients with and without severe disease in the model generation cohort (February–November 2020).

	Non Severity (n = 1548)	Severity (n = 699)	p-Value
Age (years)	66 (51–81)	83 (71–89)	<0.001
Oxygen saturation (%)	96 (94–97)	94 (91–96)	<0.001
Intermittent claudication (yes)	41 (2.65%)	42 (6.01%)	<0.001
Cerebrovascular disease (yes)	104 (6.72%)	104 (14.88%)	<0.001
Dementia (yes)	124 (8.01%)	141 (20.17%)	<0.001
Obesity (yes)	218 (14.08%)	123 (17.6%)	0.03701
Chloride (mmol/L)	101.5 (98.4–104)	102.3 (99–106)	<0.001
Creatinine (mg/dL)	0.86 (0.68–1.1)	1.1 (0.81–1.58)	<0.001
Eosinophils (%)	0.17 (0–0.7)	0 (0–0.2)	<0.001
Eosinophils (mil/mm^3)	0.0102 (0–0.04385)	0 (0–0.01702)	<0.001
Glucose (mg/dL)	110 (96–133)	128 (104–170)	<0.001
International normalized ratio-prothrombin time (INR-PT)	1.1 (1.02–1.17)	1.16 (1.06–1.305)	<0.001
Lactate dehydrogenase (U/L)	275 (219–350)	336.5 (246.8–470)	<0.001
Lymphocytes (%)	17.9 (11.65–25.98)	10.74 (6.275–18)	<0.001
Lymphocytes (mil/mm^3)	1.0875 (0.7561–1.4941)	0.8055 (0.5605–1.1458)	<0.001
Monocytes (%)	8.07 (6–10.438)	6.2 (4–8.9)	<0.001
Neutrophils (%)	72 (63.28–80.41)	81.4 (72.75–88.5)	<0.001
Red blood cells (mil/mm^3)	4.6 (4.19–4.95)	4.32 (3.87–4.72)	<0.001
Urea (g/l)	0.3595 (0.27–0.5258)	0.609 (0.42–0.91)	<0.001
Mean corpuscular volume (fl)	89.7 (86.1–93)	91.5 (87.7–95)	<0.001

Data are presented as n (%) or median (interquartile range).

Table 2. Demographic and clinical information of patients with and without severe disease in the external validation cohort (November 2020–January 2021).

	Non Severity (n = 425)	Severity (n = 201)	p-Value
Age (years)	71 (58–83)	84 (74–89)	<0.001
Oxygen saturation (%)	95 (94–97)	95 (92–97)	0.03714
Intermittent claudication (yes)	24 (5.65%)	17 (8.46%)	0.2484
Cerebrovascular disease (yes)	44 (10.35%)	37 (18.41%)	0.007451
Dementia (yes)	43 (10.12%)	32 (15.92%)	0.05051
Obesity (yes)	75 (17.65%)	27 (13.43%)	0.2236
Chloride (mmol/L)	101.7 (98.9–104.1)	102.1 (98.3–105.7)	0.1273
Creatinine (mg/dL)	0.85 (0.68–1.06)	1.14 (0.8–1.64)	<0.001
Eosinophils (%)	0.1 (0–0.4)	0.01 (0–0.24)	<0.001
Eosinophils (mil/mm^3)	0.00715 (0–0.02808)	0.00026 (0–0.0162)	<0.001
Glucose (mg/dL)	116 (100–145)	130 (106–174)	<0.001
International normalized ratio-prothrombin time (INR-PT)	1.08 (1.02–1.17)	1.15 (1.06–1.32)	<0.001
Lactate dehydrogenase (U/L)	266 (213–336)	322 (245–414)	<0.001
Lymphocytes (%)	16 (10.5–24.4)	9.2 (5.5–14.8)	<0.001
Lymphocytes (mil/mm^3)	0.9432 (0.6696–1.362)	0.6987 (0.4611–1.0577)	<0.001
Monocytes (%)	7.9 (5.6–10.59)	5.7 (3.6–8.29)	<0.001
Neutrophils (%)	74.7 (64.66–82.4)	84.5 (75.1–88.7)	<0.001
Red blood cells (mil/mm^3)	4.46 (4.01–4.81)	4.17 (3.66–4.62)	<0.001
Urea (g/L)	0.396 (0.306–0.553)	0.672 (0.446–1.1)	<0.001
Mean corpuscular volume (fl)	90.1 (87.1–93)	90.7 (87–94.9)	0.07457

Regarding the modelling procedure, Table 3 contains the set of hyperparameters explored in training the different algorithms (MLP, random forest, and XGBoost), as well as their best combination of hyperparameters and the AUC achieved. In the case of MLP, the Early Stopping technique was applied to avoid overfitting. To enhance training, both classes were weighted to balance the training sample.

Table 3. Search space of the hyperparameters explored for each algorithm.

Model	Parameters	Search Space	Best Model	AUC
MLP	Number of hidden layers	[2, 10]	7	0.8015
	Number of neurons	[16, 512]	[96, 176, 240, 240, 352, 352, 384]	
	Activation layer	[selu, linear, tanh, softmax]	[softmax, selu, softmax, selu, selu, selu, selu]	
	Learning rate	{0.001,0.01,0.1}	0.001	
	Optimizer	{sgd, adam, rmsprop}	rmsprop	
	Batch size	[1, 64]	54	
Random Forest	Number of estimators	[30, 1300]	820	0.8297
	Max. depth	[3, 20]	15	
	Min. samples split	[2, 30]	10	
	Criterion	{gini, entropy}	gini	
	Min. impurity decrease	$\{5 \times 10^{-5}, 1 \times 10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}, 1.5 \times 10^{-3}, 2 \times 10^{-3}, 5 \times 10^{-3}, 0.01\}$	2×10^{-4}	
XGBoost	Number of estimators	[30, 1300]	330	0.8307
	Scale pos. weight	[1, 10]	1	
	Column subsample size per tree	[0.3, 1]	0.85	
	Subsample size per tree	[0.3, 1]	0.64	
	Max. depth	[3, 20]	15	
	Learning rate	$\{10^{-4}, 10^{-3}, 10^{-2}$,0.1, 0.15, 0.2, 0.3, 0.4}	0.01	
	Reg. alpha	$\{10^{-4}, 10^{-3}, 10^{-2}$,0.1, 0.15, 0.2, 0.3, 0.4}	0.4	
	Gamma	[0.05, 1]	0.6	

Sgd: stochastic gradient descendent; Max. depth: The maximum tree depth for the base learners; Min. samples split: Minimum number of samples remaining in a node to consider splitting it; Min. impurity decrease: A node will be split if this split induces a decrease in the impurity greater than or equal to this value; Scale pos. weight: The balance between positive and negative classes; Reg. alpha: The L1 regularization on the weights; Gamma: The minimum loss reduction required to further partition a leaf node of the tree. Column subsample size per tree and Subsample size per tree describe the ratio of columns and rows, respectively, used in each boosting round, and Learning rate is the boosting learning rate.

The ROC curves of the best models for each ML algorithm on the test dataset are shown in Figure 4. The AUCs obtained by the algorithms that combine decision trees (i.e., random forest and XGBoost) were superior to those obtained by the MLP, with XGBoost achieving the best performance (AUC = 0.8307). The curve comparison test showed a significant predictive improvement (p = 0.043) by considering the XGBoost model vs. MLP.

For the best algorithm (XGBoost), Table 4 shows the set of hyperparameters for which the best models were attained considering all variables, the 50 best variables, or the 20 best variables as candidate variables. Figure 5 is an importance diagram of the 20 variables that were finally selected.

Table 4. Hyperparameter configuration of the best XGBoost models.

Parameter	All-Variable Model	50-Variable Model	20-Variable Model
Number of estimators	330	690	340
Scale pos. weight	1	5	1
Column subsample size per tree	0.85	0.41	0.84
Subsample size per tree	0.64	0.94	0.68
Max. depth	15	14	13
Learning rate	0.01	0.2	0.01
Reg. alpha	0.4	0.3	0.15
Gamma	0.6	0.1	0.45

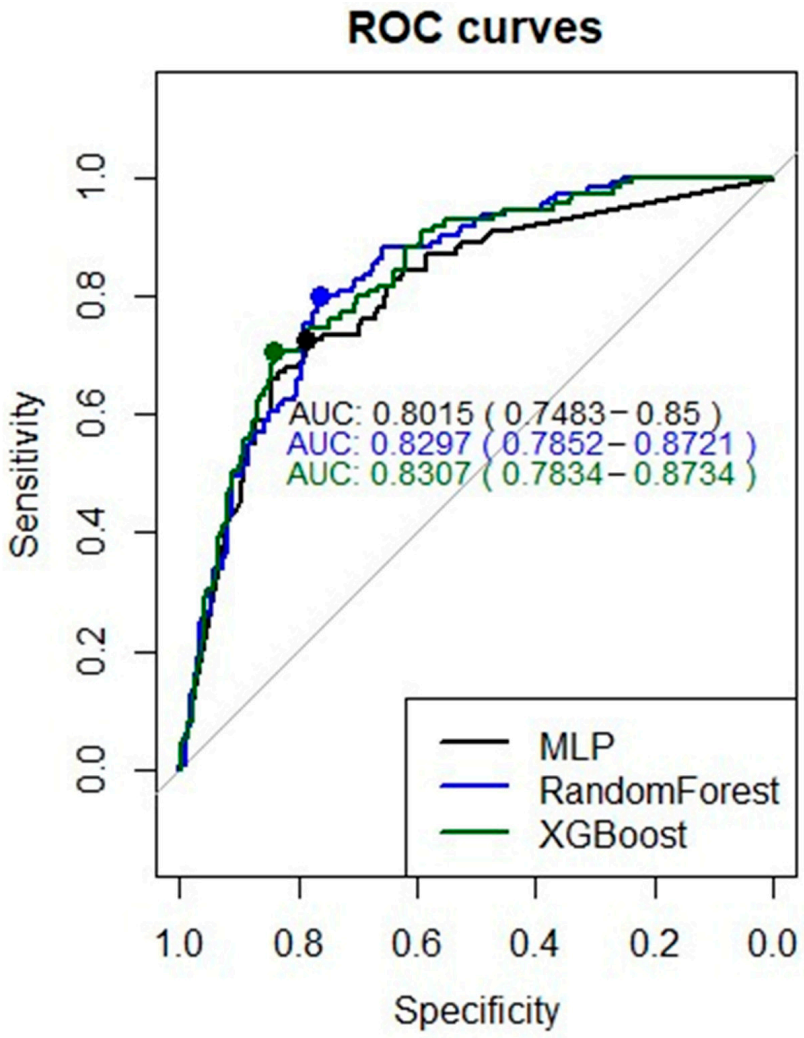


Figure 4. ROC curves of the best models (MLP, Random Forest, and XGBoost).

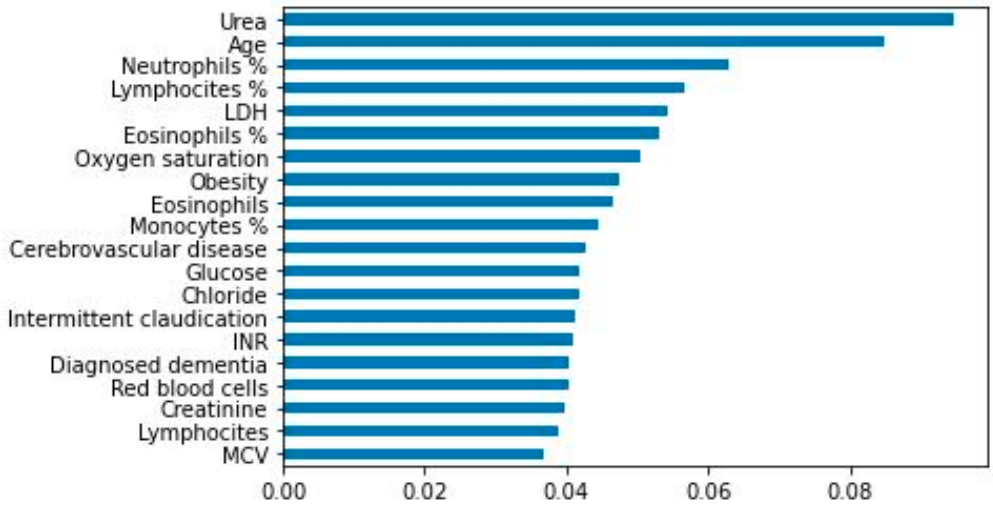


Figure 5. Importance diagram of the 20 variables selected.

Finally, the ROC curves for the final models analysing all variables, 50 variables, and 20 variables on the test dataset are shown in Figure 6. The AUC comparison revealed non-significant predictive improvement or loss ($p > 0.1$) when considering the best model with the 165 source variables (AUC = 0.8307; 95% CI 0.7856–0.8718), the best model with

the 50 most important variables (AUC = 0.8138; 95% CI 0.7654–0.8571), and the best model with the 20 most important variables (AUC = 0.8153; 95% CI 0.7655–0.8615). Therefore, the most robust and parsimonious choice was to consider the 20-variable model as the final model due to its simplicity without loss of predictive capacity.

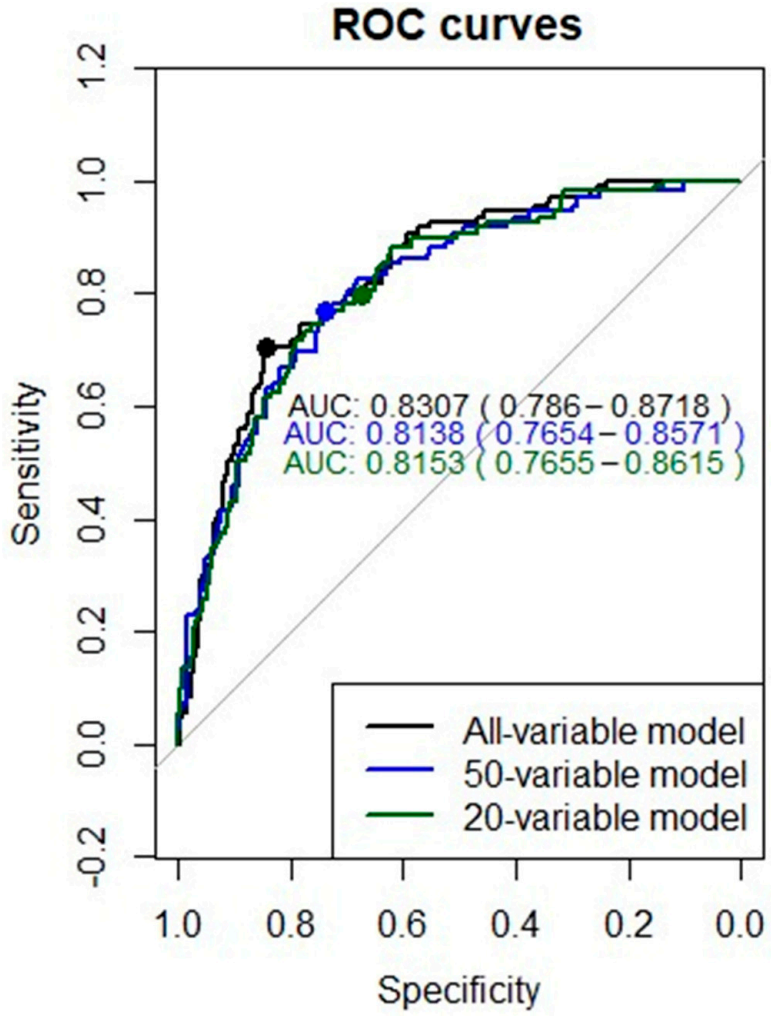


Figure 6. ROC curves of the analysed models considering all variables, 50 variables, and the 20 most influential variables.

3.1. External Validation Analysis

Regarding the model validation, Figure 7 shows the boxplots for each real class (non-severity; severity, respectively) on the external validation dataset. The distribution of the probabilities predicted by the model for non-severity patients (green box) was clearly under the distribution of probabilities for severity patients in the red box. Thus, we observed good discrimination ability for the predictive model and a possible threshold of 0.4 that can separate the two groups.

Figure 8 shows good agreement between the predicted probabilities and real outcome in our external validation, with an intercept of −0.123 and a slope of 1.006. Moreover, the ROC curve (Figure 9, AUC = 0.821; 95% CI 0.787–0.854) demonstrates no loss in predictive ability in the external validation. The Youden index (4) was obtained with a specificity (2) and sensitivity (3) pair of 0.609 and 0.886, respectively.

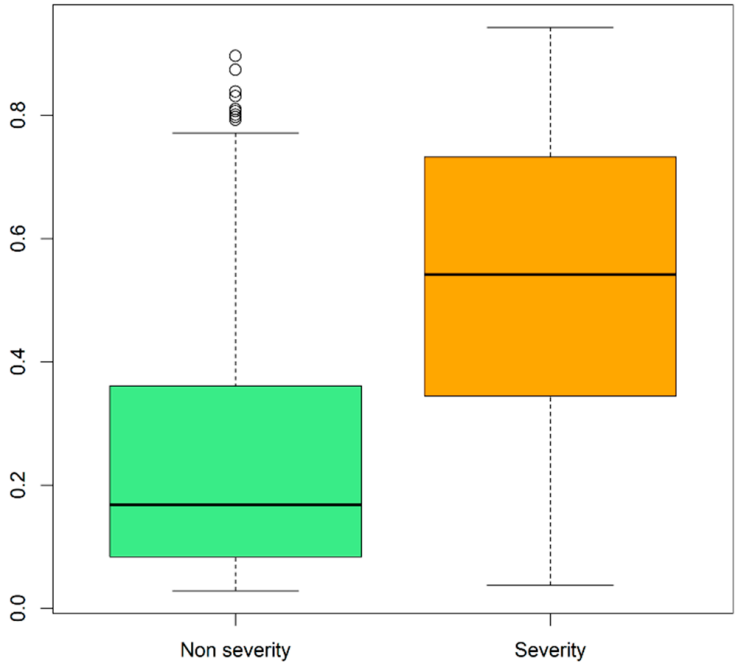


Figure 7. Boxplots of the probabilities predicted by the final model for the external validation set grouped by the actual class group.

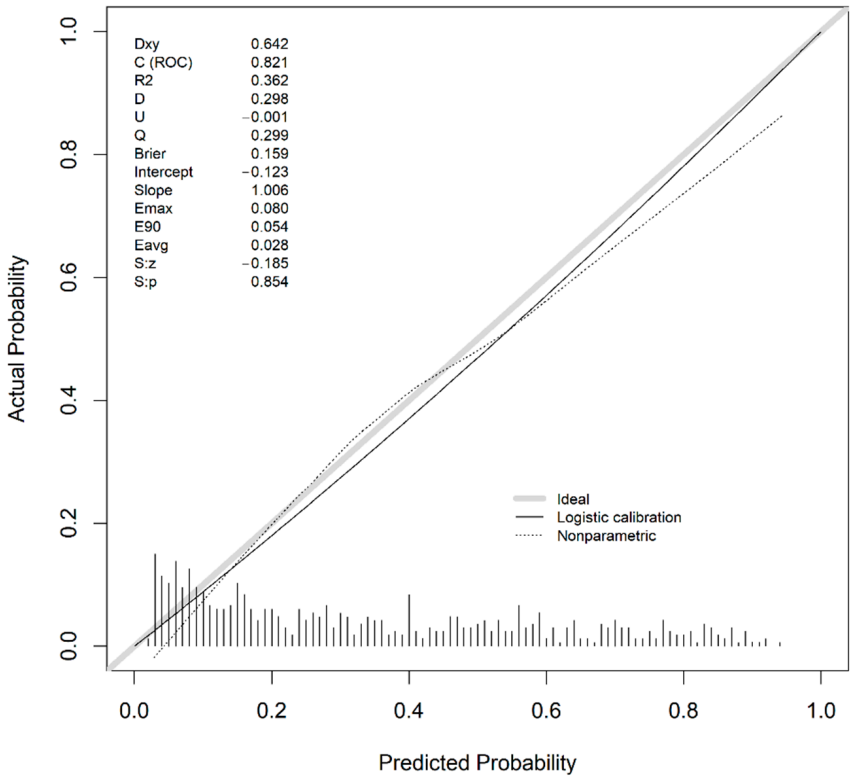


Figure 8. Calibration curve for the external validation dataset.

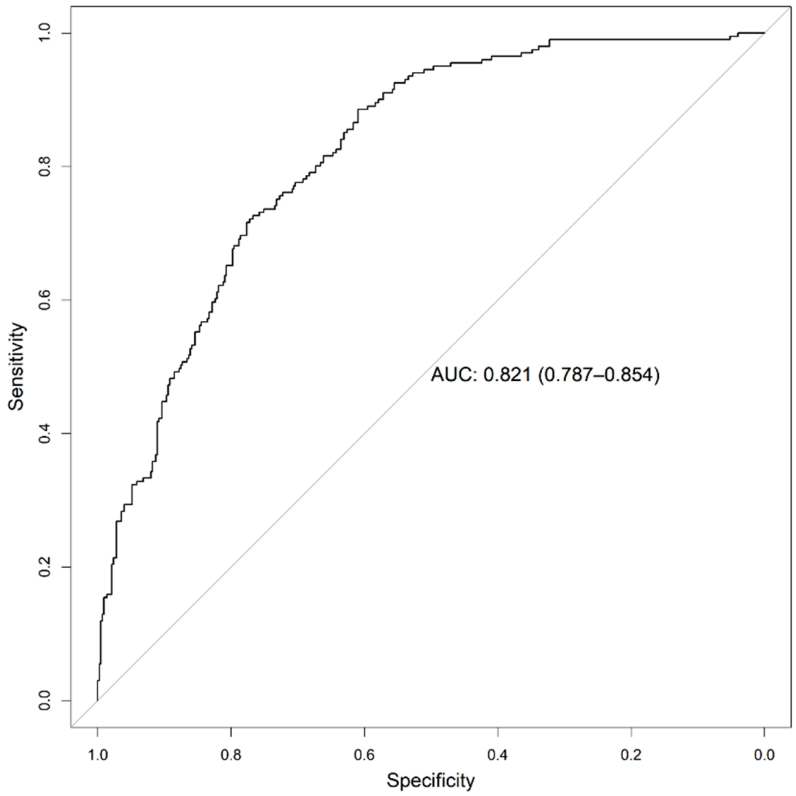


Figure 9. ROC curve for the external validation dataset.

The clinical utility analysis of the model is given in Table 5. For each probability threshold, the following metrics are presented: sensitivity (3), specificity (2), PPV (5), NPV (6), and accuracy (1). The cut-offs of 0.24 and 0.4 equally optimize the Youden index (4). However, if minimization of both classification errors is desired, a 0.4 threshold gives us more balanced (0.71 and 0.78) sensitivity (3) and specificity (2) values.

Table 5. Cut-off analysis of the external validation cohort (November 2020–January 2021).

Thr	tp	tn	fp	fn	Sens	Spec	PPV	NPV	Accuracy	Youden
0.05	199	44	381	2	0.99	0.1	0.34	0.96	0.39	0.09
0.1	199	137	288	2	0.99	0.32	0.41	0.99	0.54	0.31
0.15	192	187	238	9	0.96	0.44	0.45	0.95	0.61	0.4
0.2	184	237	188	17	0.92	0.56	0.49	0.93	0.67	0.47
0.22	179	250	175	22	0.89	0.59	0.51	0.92	0.69	0.48
0.24	178	257	168	23	0.89	0.6	0.51	0.92	0.69	0.49
0.26	171	267	158	30	0.85	0.63	0.52	0.9	0.7	0.48
0.28	164	277	148	37	0.82	0.65	0.53	0.88	0.7	0.47
0.3	159	288	137	42	0.79	0.68	0.54	0.87	0.71	0.47
0.32	154	300	125	47	0.77	0.71	0.55	0.86	0.73	0.47
0.34	152	307	118	49	0.76	0.72	0.56	0.86	0.73	0.48
0.36	148	318	107	53	0.74	0.75	0.58	0.86	0.74	0.48
0.38	146	326	99	55	0.73	0.77	0.6	0.86	0.75	0.49
0.4	143	330	95	58	0.71	0.78	0.6	0.85	0.76	0.49
0.42	134	339	86	67	0.67	0.8	0.61	0.83	0.76	0.46
0.44	131	343	82	70	0.65	0.81	0.62	0.83	0.76	0.46
0.46	125	345	80	76	0.62	0.81	0.61	0.82	0.75	0.43
0.48	117	353	72	84	0.58	0.83	0.62	0.81	0.75	0.41
0.5	113	359	66	88	0.56	0.84	0.63	0.8	0.75	0.41

Thr: Probability threshold, tp: True positive, tn: True negative, fp: False positive, fn: False negative, Sens: Sensitivity = tp/(tp + fn), Spec: Specificity = tn/(tn + fp), PPV: Positive predictive value = tp/(tp + fp), NPV: Negative predictive value = tn/(tn + fn), Accuracy = (tp + tn)/(tp + fp + tn + fn), Youden = Sens + Spec − 1 = tp/(tp + fn) + tn/(tn + fp) − 1.

3.2. Clinically Useful Tool

The generated model could be used in clinical practice via a user-friendly web application. The necessary code to assemble the tool and apply the generated model is available at https://github.com/ITAINNOVA/covid_IIS (accessed on 6 July 2021).

The designed tool has the main functionality of predicting severity (death or ICU admission) through the best performing model given the information for the 20 explanatory variables, which are the fields to be entered into the tool. Given the patient information, the tool returns the probability of the patient ending up in a serious condition (i.e., ICU or death). Although the number of variables to be entered into the tool is not high and is consistent with practical use, the information for some of the variables may not be available for some patients. In these cases, we would encounter a problem of incompleteness, which we solve by assigning the mean value.

In addition to providing the risk given the patient's characteristics, the interpretability and explainability of the model were explored to provide more information to the clinician. In particular, using the Shapley technique [50], the tool provides a graphical representation of the Shapley values using Python's SHAP library [51], which allows interpretation of the direction and intensity of each variable. An example of such a representation provided by the tool for a given patient is shown in Figure 10; the patient's age (lower than average) contributes the most (has the longest bar) to decreasing the probability of risk, but the amount of urea in the patient contributes to the contrary.

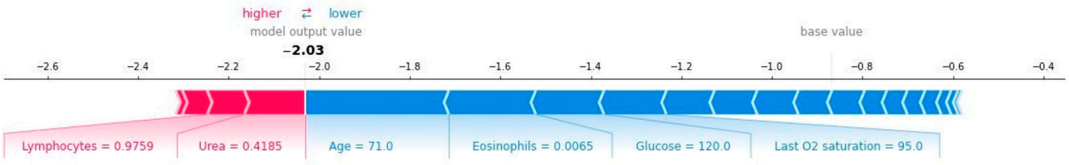


Figure 10. Example of a graphical representation of the Shapley values provided by the tool. *Base value*: mean prediction of the test set. Blue variables contribute to decreasing the probability of the patient risk, whereas red variables contribute to increasing the probability of patient risk. The magnitude of such increase or decrease is measured by the length of the bar.

4. Discussion

A large number of studies have analysed predictive risk factors for COVID-19 [52–54], though less common are studies that have explored predictive models. Our study focused on the generation of a risk prediction model using machine learning techniques and variables that are easily obtained in the emergency room of most hospitals.

We trained neural networks, random forest and XGBoost algorithms using a hyperparameter tuning optimization. Our best model was reached using XGBoost algorithm, that taking into account the amount of candidate predictors can be a good alternative. This kind of models provides robust models as their predictions are based on an additive combination of trees that are built using different sets of data and variables. In our case, the best model was reached using 330 trees, those trees are built using the 85% of predictor variables and 64% of the training data sample, this guarantees that each tree explored the predictive ability of predictor variables in different data sample and over a different set of variables. In addition, the trees had a maximum depth of 15, preventing the overfitting that it is present in trees with too many branches.

The model we developed was built from a cohort of patients in Aragon who were hospitalized with a positive SARS-CoV-2 PCR test, and the outcome studied was ICU admission or mortality within 30 days of hospital admission. The analyses were performed for a cohort of 3623 patients, which is larger than the predictive models reviewed by Wynants et al. [23]. Several studies have analysed predictive models of mortality, but with an unclear definition of the target population, without specifying the outcome (mortality) period, and at high risk of biases [33–36]. This bias can lead to miscalibrations and overestimation of the discrimination performance, especially when they are not externally

validated. Our model was validated both internally and externally to avoid overfitting of the generated model and to provide a true measure of model performance and generalizability. The internal validation was based on 12.5% of the training data and provides a good performance AUC = 0.8307 of our model. The set used for external validation maintained a proportion of patients (n = 626) with outcomes (death or ICU) similar to the training set (31–32%). The AUC value of 0.821 shown a minimum loss in the predictive ability of the model.

Regarding calibration, Xie et al. [39], despite achieving excellent discrimination (C index = 0.98), obtained a calibration that had a slope > 1 (probabilities too high for low-risk patients and too low for high-risk patients). Our model achieved good discrimination with the external validation set (AUC = 0.821, 95% CI 0.787–0.854) and accurate calibration (slope = 1, intercept = −0.12) [55], which means that the predicted probabilities are close to the expected probability distribution.

Most predictive models usually show their accuracy by the AUC as a measure of the discrimination ability. The Marcos et al. [28] model achieved an AUC of 0.83 (95% CI 0.81–0.85). With a similar aim, models trained by Patel et al. [30] with only the top five features achieved similar performance as those using all features; in particular, the model for ICU admission achieved an AUC of 0.79 (95% CI 0.72–0.85) and the one for mechanical ventilation achieved an AUC of 0.83 (95% CI 0.77–0.9). Knight et al. [23] evaluated the discrimination of the model, which obtained an AUC of 0.77 (95% CI 0.76–0.77), and its calibration, which was excellent (calibration-in-the-large = 0, slope = 1.0). Therefore, our model achieves a performance close to or even superior to that of the literature studies.

The study we present explored machine learning techniques for the development and validation of models with the aim of predicting an unfavourable outcome for a patient admitted with COVID-19, where an unfavourable outcome is understood to be ICU admission or mortality. In particular, MLP, random forest and XGBoost were analysed using dynamic optimization of the model's hyperparameters. The ultimate goal of our study was to build a model that can be used in clinical practice via a user-friendly web-based tool/platform accessible by the clinician. To this end, the 20 most important variables readily available in the emergency and patient records were selected for generation of the model. The final model was built using the XGBoost algorithm and achieved an AUC of 0.821 (95% CI 0.787–0.854) in the external validation set.

In contrast to the reviewed papers, our study included a large cohort with a great amount of data. In addition, as discussed above, our study was validated both internally and externally, showing good calibration, on a set of patients that includes two different pandemic waves in the training set and a third one in the validation set, but maintaining a similar percentage of severity (unfavourable outcomes).

An analysis of the optimal cut-off points was also carried out depending on the metric to be optimized, which provides a series of thresholds for decision-making, allowing one cut-off or another to be chosen depending on the needs of the system. The probability threshold of 40% provided the best performance in a sensitivity/specificity equilibrium analysis, but we provided a clinical utility analysis on different cut-offs to choose another pair of sensitivity (3) and specificity (2) values depending on the evolution of the pandemic.

Although the generated model was externally validated with good calibration and discrimination ability, the study has some limitations. Our study was carried out with data from a single region of Spain (Aragon), but the hospitals belong to the National Health System, which has universal health coverage and, therefore, contains information from patients of different ethnicities and social characteristics. For this reason, diversity is guaranteed. Although the model considers patient variables that have been shown to be risk factors, we propose also considering the inclusion of other variables of interest in future studies, such as the type of vaccine administered or the time of vaccination. Regarding model limitations, boosting techniques are more likely to overfit than bagging although we have used hyperparameter tuning to prevent this overfitting. We also encourage validation of our model with a different cohort.

5. Conclusions

In this study, we incorporated information from a population that includes several pandemic waves to generate and validate a model capable of predicting death or admission to the ICU for patients admitted with COVID-19 based on 20 emergency and clinical history characteristics. The results of the internal and external validation showed a good discriminative capacity and calibration of the model, which implies that the use of the model is adequate. The final objective of the study was to design and provide an ergonomic web application (https://github.com/ITAINNOVA/covid_IIS) (accessed on 6 July 2021) that integrates the model in a way that provides the clinician with the probability that the patient admitted for COVID ends up serious, which can aid in developing an action plan.

Author Contributions: Conceptualization, J.R.P.-P., M.J.E.-R., Á.L. and M.T.S.; methodology, R.A.-G., L.M.E., G.L.-L. and R.d.-H.-A.; software, R.A.-G., G.L.-L. and D.A.-G.; validation, R.A.-G., L.M.E., G.L.-L., R.d.-H.-A., Á.L. and M.T.S.; formal analysis, R.A.-G., L.M.E., G.L.-L. and R.d.-H.-A.; investigation, R.A.-G., L.M.E., G.L.-L., R.d.-H.-A. and M.T.S.; resources, R.A.-G., L.M.E., G.L.-L., D.A.-G. and R.d.-H.-A.; data curation, R.A.-G. and G.L.-L.; writing—original draft preparation, R.A.-G., L.M.E., G.L.-L., R.d.-H.-A. and M.T.S., writing—review and editing, R.A.-G., L.M.E., G.L.-L., R.d.-H.-A. and M.T.S.; visualization, R.A.-G., L.M.E., G.L.-L. and R.d.-H.-A., supervision, J.R.P.-P., M.J.E.-R., Á.L. and M.T.S.; project administration, R.d.-H.-A. and M.T.S.; funding acquisition, R.d.-H.-A. and Á.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Health Research Institute of Aragon (IIS Aragon) and ITAINNOVA-Instituto Tecnológico de Aragón.

Institutional Review Board Statement: The research protocol was approved by the Clinical Research Ethics Committee of Aragon (PI 20/183).

Informed Consent Statement: Due to the retrospective, observational nature of this study, the data could be fully anonymized and informed consent was waived.

Data Availability Statement: Data sharing not applicable.

Acknowledgments: The authors would like to thank their patients and fellow health care workers for providing excellent medical care at considerable personal risk.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

AUC	Area Under ROC Curve
CI	Confidence Interval
COVID-19	Coronavirus disease 2019
ICU	Intensive Care Unit
INR	International Normalized Ratio
INR-PT	International Normalized Ratio-Prothrombin Time
LDH	Lactate Dehydrogenase
MCV	Mean Corpuscular Volume
ML	Machine Learning
MLP	Multilayer Perceptron
NPV	Negative Predicted Value
PCR	Polymerase Chain Reaction
PPV	Positive Predictive Value
ROC	Receiver Operating Characteristic
RT-PCR	Reverse Transcription Polymerase Chain Reaction
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SVM	Support Vector Machine
TPE	Tree-structured Parzen Estimator
XGBoost	Extreme Gradient Boosting

References

1. DeLancey, E.R.; Simms, J.F.; Mahdianpari, M.; Brisco, B.; Mahoney, C.; Kariyeva, J. Comparing deep learning and shallow learning for large-scale wetland classification in Alberta, Canada. *Remote Sens.* **2020**, *12*, 2. [CrossRef]
2. Thanh, H.V.; Sugai, Y.; Sasaki, K. Application of artificial neural network for predicting the performance of CO_2 enhanced oil recovery and storage in residual oil zones. *Sci. Rep.* **2020**, *10*, 18204. [CrossRef]
3. Lacueva-Pérez, F.J.; Ilarri, S.; Vargas, J.J.B.; Lezaun, G.L.; Alonso, R.D.H. Multifactorial Evolutionary Prediction of Phenology and Pests: Can Machine Learning Help? In Proceedings of the WEBIST 2020—16th International Conference on Web Information Systems and Technologies, Online, 3–5 November 2020; Science and Technology Publications, Lda: Setúbal, Portugal, 2020; pp. 75–82.
4. Han, T.; Li, Y.F.; Qian, M. A hybrid generalization network for intelligent fault diagnosis of rotating machinery under unseen working conditions. *IEEE Instrum. Meas.* **2021**, *70*, 3520011. [CrossRef]
5. Khan, S.; Yairi, T. A review on the application of deep learning in system health management. *Mech. Syst. Signal. Process.* **2018**, *107*, 241–265. [CrossRef]
6. Jiang, L.; Wu, Z.; Xu, X.; Zhan, Y.; Jin, X.; Wang, L.; Qiu, Y. Opportunities and challenges of artificial intelligence in the medical field: Current application, emerging problems, and problem-solving strategies. *J. Int. Med. Res.* **2021**, *49*, 03000605211000157. [CrossRef] [PubMed]
7. Secinaro, S.; Calandra, D.; Secinaro, A.; Muthurangu, V.; Biancone, B. The role of artificial intelligence in healthcare: A structured literature review. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 125. [CrossRef]
8. Musulin, J.; Baressi Šegota, S.; Štifanić, D.; Lorencin, I.; Anđelić, N.; Šušteršič, T.; Blagojević, A.; Filipović, A.; Ćabov, T.; Markova-Car, E. Application of Artificial Intelligence-Based Regression Methods in the Problem of COVID-19 Spread Prediction: A Systematic Review. *Int. J. Environ. Res. Public Health* **2021**, *18*, 4287. [CrossRef]
9. Aznar-Gimeno, R.; Labata-Lezaun, G.; Adell-Lamora, A.; Abadía-Gallego, D.; del-Hoyo-Alonso, R.; González-Muñoz, C. Deep Learning for Walking Behaviour Detection in Elderly People Using Smart Footwear. *Entropy* **2021**, *23*, 777. [CrossRef]
10. Guo, Y.; Liu, Y.; Oerlemans, A.; Lao, S.; Wu, S.; Lew, M.S. Deep learning for visual understanding: A review. *Neurocomputing* **2016**, *187*, 27–48. [CrossRef]
11. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep learning for computer vision: A brief review. *Comput. Intell. Neurosci.* **2018**, *2018*, 7068349. [CrossRef] [PubMed]
12. Kim, G.B.; Jung, K.H.; Lee, Y.; Kim, H.-J.; Kim, N.; Jun, S.; Seo, J.B.; Lynch, D.A. Comparison of shallow and deep learning methods on classifying the regional pattern of diffuse lung disease. *J. Digit. Imaging* **2018**, *31*, 415–424. [CrossRef]
13. Chauhan, S.; Vig, L.; De Filippo De Grazia, M.; Corbetta, M.; Ahmad, S.; Zorzi, M. A comparison of shallow and deep learning methods for predicting cognitive performance of stroke patients from MRI lesion images. *Front. Neuroinform.* **2019**, *13*, 53. [CrossRef] [PubMed]
14. Del CarmenRodríguez-Hernández, M.; del-Hoyo-Alonso, R.; Ilarri, S.; Montafñés-Salas, R.M.; Sabroso-Lasa, S. An Experimental Evaluation of Content-based Recommendation Systems: Can Linked Data and BERT Help? In Proceedings of the 2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA), Antalya, Turkey, 2–5 November 2020; IEEE: Antalya, Turkey, 2020; pp. 1–8.
15. Emmert-Streib, F.; Yang, Z.; Feng, H.; Tripathi, S.; Dehmer, D. An introductory review of deep learning for prediction models with big data. *Front. Artif. Intell.* **2020**, *3*, 4. [CrossRef]
16. Playe, B.; Stoven, V. Evaluation of deep and shallow learning methods in chemogenomics for the prediction of drugs specificity. *J. Cheminform.* **2020**, *12*, 1–18. [CrossRef] [PubMed]
17. Chen, J.; See, K.C. Artificial Intelligence for COVID-19: Rapid Review. *J. Med. Internet Res.* **2020**, *22*, e21476. [CrossRef]
18. Global Data: Coronavirus Pandemic COVID-19. Available online: https://www.worldometers.info/coronavirus/ (accessed on 9 June 2021).
19. Wu, Z.; McGoogan, J.M. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: Summary of a report of 72,314 cases from the Chinese Center for Disease Control and Prevention. *JAMA* **2020**, *323*, 1239–1242. [CrossRef]
20. Karadag, E. Increase in COVID-19 cases and case-fatality and case-recovery rates in Europe: A cross-temporal meta-analysis. *J. Med. Virol.* **2020**, *92*, 1511–1517. [CrossRef]
21. Izcovich, A.; Ragusa, M.A.; Tortosa, F.; Lavena Marzio, M.A.; Agnoletti, C.; Bengolea, A.; Ceirano, A.; Espinosa, F.; Saavedra, E.; Sanguine, V. Prognostic factors for severity and mortality in patients infected with COVID-19: A systematic review. *PLoS ONE* **2020**, *15*, e0241955.
22. Knight, S.R.; Ho, A.; Pius, R.; Buchan, I.; Carson, G.; Drake, T.M.; Dunning, J.; Fairfield, C.J.; Gamble, C.; Green, C.A. Risk stratification of patients admitted to hospital with COVID-19 using the ISARIC WHO Clinical Characterisation Protocol: Development and validation of the 4C Mortality Score. *BMJ* **2020**, *370*, m3339. [CrossRef]
23. Wynants, L.; Van Calster, B.; Collins, G.S.; Riley, R.D.; Heinze, G.; Schuit, E.; Bonten, M.M.J.; Dahly, D.L.; Damen, J.A.; Debray, T.P.A. Prediction models for diagnosis and prognosis of COVID-19 infection: Systematic review and critical appraisal. *BMJ* **2020**, *369*, m1328. [CrossRef] [PubMed]

24. Aznar-Gimeno, R.; Paño-Pardo, J.R.; Esteban, L.M.; Labata-Lezaun, G.; Esquillor-Rodrigo, M.J.; Lanas, A.; Abadía-Gallego, D.; Diez-Fuertes, F.; Tellería-Orriols, C.; del-Hoyo-Alonso, R.; et al. Changes and Evolution among SARS-COV-2 Hospitalised Patients in Terms of Severity, Mortality and Virus Genome in a Spanish Cohort. Research Square [Preprint] (2021). Available online: https://www.researchsquare.com/article/rs-199395/v1 (accessed on 9 June 2021).
25. Cai, W.; Liu, T.; Xue, X.; Luo, G.; Wang, X.; Shen, Y.; Fang, Q.; Sheng, J.; Dhen, F.; Liang, T. CT Quantification and Machine-learning Models for Assessment of Disease Severity and Prognosis of COVID-19 Patients. *Acad. Radiol.* **2020**, *27*, 1665–1678. [CrossRef] [PubMed]
26. Wu, G.; Yang, P.; Xie, Y.; Woodruff, H.C.; Rao, X.; Guiot, J.; Frix, A.N.; Louis, R.; Moutschen, M.; Li, J.; et al. Development of a clinical decision support system for severity risk prediction and triage of COVID-19 patients at hospital admission: An international multicentre study. *Eur. Respir. J.* **2020**, *56*, 2001104. [CrossRef]
27. Yao, H.; Zhang, N.; Zhang, R.; Duan, M.; Xie, T.; Pan, J.; Peng, E.; Huang, J.; Zhang, Y.; Xu, X.; et al. Severity detection for the coronavirus disease 2019 (COVID-19) patients using a machine learning model based on the blood and urine tests. *Front. Cell Dev. Biol.* **2020**, *8*, 683. [CrossRef]
28. Marcos, M.; Belhassen-Garcia, M.; Sanchez-Puente, A.; Sampedro-Gomez, J.; Azibeiro, R.; Dorado-Díaz, P.I.; Marcano-Millar, E.; García-Vidal, C.; Moreiro-Barroso, M.T.; Cubino-Bóveda, N.; et al. Development of a severity of disease score and classification model by machine learning for hospitalized COVID-19 patients. *PLoS ONE* **2020**, *16*, e0240200.
29. Jiang, X.; Coffee, M.; Bari, A.; Wang, J.; Jiang, X.; Huang, J.; Shi, J.; Dai, J.; Cai, J.; Zhang, T.; et al. Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. *CMC Comput. Mater. Con.* **2020**, *63*, 537–551. [CrossRef]
30. Patel, D.; Kher, V.; Desai, B.; Lei, X.; Cen, S.; Nanda, N.; Gholamrezanezhad, A.; Duddalwar, V.; Varghese, B.; Oberai, A.A. Machine learning based predictors for COVID-19 disease severity. *Sci. Rep.* **2021**, *11*, 4673. [CrossRef]
31. Štifanić, D.; Musulin, J.; Jurilj, Z.; Šegota, S.B.; Lorencin, I.; Anđelić, N.; Vlahinić, S.; Šušteršič, T.; Blagojević, A.; Filipović, N.; et al. Semantic segmentation of chest X-ray images based on the severity of COVID-19 infected patients. *EAI Endorsed Trans. Bioeng. Bioinform.* **2021**, *1*, e3.
32. Riley, R.D.; Ensor, J.; Snell, K.I.; Harrell, F.E.; Martin, G.P.; Reitsma, J.B.; Moons, K.G.M.; Collins, G.; van Smeden, M. Calculating the sample size required for developing a clinical prediction model. *BMJ* **2020**, *368*, m441. [CrossRef]
33. Caramelo, F.; Ferreira, N.; Oliveiros, B. Estimation of risk factors for COVID-19 mortality-preliminary results. *medRxiv* **2020**. [CrossRef]
34. Yan, L.; Zhang, H.T.; Xiao, Y.; Wang, M.; Guo, Y.; Sun, C.; Tang, X.; Jing, L.; Li, S.; Zhang, M.; et al. Prediction of criticality in patients with severe COVID-19 infection using three clinical features: A machine learning-based prognostic model with clinical data in Wuhan. *medRxiv* **2020**. [CrossRef]
35. Yuan, M.; Yin, W.; Tao, Z. Association of radiologic findings with mortality of patients infected with 2019 novel coronavirus in Wuhan, China. *PLoS ONE* **2020**, *15*, e0230548. [CrossRef]
36. Shi, Y.; Yu, X.; Zhao, H. Host susceptibility to severe COVID-19 and establishment of a host risk score: Findings of 487 cases outside Wuhan. *Crit. Care* **2020**, *24*, 108. [CrossRef]
37. Yue, H.; Yu, Q.; Liu, C.; Huang, Y.; Jiang, Z.; Shao, C.; Zhang, H.; Ma, B.; Wang, Y.; Xie, G.; et al. Machine learning-based CT radiomics model for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: A multicenter study. *Ann. Transl. Med.* **2020**, *8*, 859. [CrossRef] [PubMed]
38. Gong, J.; Ou, J.; Qiu, X.; Jie, Y.; Chen, Y.; Yuan, L.; Cao, J.; Tan, M.; Xu, W.; Zheng, F.; et al. A tool to early predict severe 2019-novel coronavirus pneumonia (COVID-19): A multicenter study using the risk nomogram in Wuhan and Guangdong, China. *Clin. Infect. Dis.* **2020**, *71*, 833–840. [CrossRef] [PubMed]
39. Xie, J.; Hungerford, D.; Chen, H.; Abrams, S.T.; Li, S.; Wang, G.; Wang, Y.; Kang, H.; Bonnett, L.; Zheng, R.; et al. Development and External Validation of a Prognostic Multivariable Model on Admission for Hospitalized Patients with COVID-19. 2020. Available online: https://www.medrxiv.org/content/medrxiv/early/2020/03/30/2020.03.28.20045997.full.pdf (accessed on 14 July 2021).
40. Koller, D.; Sahami, M. Toward optimal feature selection. In Proceedings of the ICML'96 Proceedings of the 13th International Conference on International Conference on Machine Learning, Bari, Italy, 3–6 July 1996.
41. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA, 2016; pp. 785–794.
42. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2009.
43. DeLong, E.R.; DeLong, D.M.; Clarke-Pearson, D.L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **1988**, *44*, 837–845. [CrossRef]
44. Pepe, M.S.; Longton, G.; Janes, H. Estimation and comparison of receiver operating characteristic curves. *Stata J.* **2009**, *9*, 1–16. [CrossRef]
45. Hosmer, D.W., Jr.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*; John Wiley & Sons: New York, NY, USA, 2013; Volume 398.
46. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'19), Anchorage, AK, USA, 4–8 August 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 2623–2631.

47. Bergstra, J.; Bardenet, R.; Bengio, Y.; Kég, B. Algorithms for hyper-parameter optimization. In Proceedings of the 25th Annual Conference on neural Information Processing Systems (NIPS) 2011, Granada, Spain, 12–14 December 2011; pp. 2546–2554.
48. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020; Available online: https://www.R-project.org/ (accessed on 14 May 2021).
49. The Python Tutorial. Available online: https://docs.python.org/3/tutorial/ (accessed on 14 May 2021).
50. Strumbelj, E.; Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.* **2015**, *41*, 647–665. [CrossRef]
51. Lundberg, S.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. *arXiv* **2017**, arXiv:1705.07874.
52. Zhou, F.; Yu, T.; Du, R.; Fan, G.; Liu, Y.; Liu, Z.; Xiang, J.; Wang, Y.; Song, B.; Gu, X.; et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study. *Lancet* **2020**, *395*, 1054–1062. [CrossRef]
53. Berenguer, J.; Ryan, P.; Rodríguez-Baño, J.; Jarrín, I.; Carratalà, J.; Pachón, J.; Yllescas, M.; Arriba, J.R.; COVID-19@Spain Study Group. Characteristics and predictors of death among 4035 consecutively hospitalized patients with COVID-19 in Spain. *Clin. Microbiol. Infect.* **2020**, *26*, 1525–1536. [CrossRef]
54. Grasselli, G.; Greco, M.; Zanella, A.; Albano, G.; Antonelli, M.; Bellani, G.; Bonanomi, E.; Cabrini, L.; Carlesso, E.; Castelli, G.; et al. Risk Factors Associated with Mortality among Patients with COVID-19 in Intensive Care Units in Lombardy, Italy. *JAMA Intern. Med.* **2020**, *180*, 1345–1355. [CrossRef]
55. Steyerberg, E.W. *Clinical Prediction Models*; Springer International Publishing: Cham, Switzerland, 2019.

7.2. Desarrollo de estándares de crecimiento fetal en embarazos gemelares

En el ámbito de la obstetricia, es común utilizar tablas de percentiles basadas en estándares de crecimiento fetal para clasificar a los fetos según su peso estimado y los puntos de corte que definen los percentiles. Esto permite identificar fetos que son pequeños para su edad gestacional (SGA, siglas en inglés) y aquellos que son grandes para su edad gestacional (LGA, siglas en inglés), que se definen como aquellos cuyo peso se encuentra por debajo del percentil 10 o por encima del percentil 90, respectivamente. La correcta identificación de estos fetos es de especial interés, debido a que han sido asociados con un mayor riesgo de resultados perinatales adversos [95, 96].

Aunque se han desarrollado múltiples estándares para gestaciones de un solo feto, su uso incorrecto en gestaciones gemelares puede resultar en intervenciones no necesarias y partos prematuros, ya que se consideraría como SGA a más del 10 %. Por ello, es esencial desarrollar y aplicar estándares específicos para gestaciones gemelares que permitan identificar de manera más precisa posibles problemas perinatales, como la muerte fetal intrauterina. Además, es fundamental considerar diferentes modelos según la corionicidad placentaria, puesto que los resultados y el peso al nacer pueden variar también entre gestaciones dicoriónicas-diamnióticas y monocoriónicas-diamnióticas. El trabajo [97] que se presenta en esta sección tuvo como objetivo desarrollar estándares de crecimiento fetal para gestacionales gemelares por corionicidad placentaria en una población española. Los estándares desarrollados fueron comparados con los estándares gemelares europeos y americanos, utilizando nuestra cohorte para calcular el porcentaje de fetos clasificados como SGA y LGA según dichos estándares.

Nuestra base de datos incluía mediciones repetidas del peso estimado gestacional para cada feto y madre. Esta jerarquía viola la suposición estadística de independencia entre las observaciones. Dada esta tipología de datos, los modelos lineales mixtos ofrecen un método adecuado para modelar la relación entre la edad gestacional y el peso estimado de estas observaciones, como se ha explicado en la sección 4.3.1. En nuestro estudio, utilizamos modelos lineales mixtos para analizar la relación entre la edad gestacional y el peso estimado de estas observaciones correlacionadas, considerándolas como efectos aleatorios. En concreto, la edad gestacional se consideró un efecto fijo, mientras que se incluyeron efectos aleatorios para capturar la variabilidad entre embarazos y fetos. Específicamente, se utilizaron splines cúbicos para modelar la edad gestacional debido a su comportamiento no lineal, con una tasa de crecimiento mayor en el segundo trimestre y más baja en el tercero. Como efectos aleatorios se incluyó un intercepto aleatorio para cada feto y una pendiente aleatoria entre el embarazo y

la estructura polinómica cúbica de la edad gestacional. Para calcular los percentiles de peso estimados, se estimó la varianza del modelo lineal en cada semana gestacional considerando los efectos aleatorios.

Nuestros modelos de crecimiento fetal mostraron un ajuste razonable a los datos, estimando aproximadamente el 10 % de los fetos de la cohorte como SGA para ambas corionicidades, lo que coincide con su definición. En comparación, los otros estándares de crecimiento fetal tendieron en su mayoría a subestimar o sobreestimar los casos de SGA y LGA. Adicionalmente, se realizó un análisis de concordancia entre los SGA y LGA proporcionados por los estándares. Finalmente, considerando como clase positiva los SGA clasificados por nuestro modelo, se calculó el valor de la sensibilidad y valor predictivo positivo como medida de capacidad de predicción del SGA, población más vulnerable. Los resultados mostraron, en general, unos valores superiores al 60 % en las métricas en el caso de los embarazos gemelares dicoriónicos. Asimismo, este trabajo de investigación ha dado lugar a la creación de la librería PTwins [2] en R, donde se ofrece el uso del modelo desarrollado. El manual en el que se detalla su uso se encuentra en el Anexo B.

En resumen, este estudio se centró en el desarrollo de modelos de crecimiento fetal para embarazos gemelares utilizando modelos lineales mixtos. En este contexto, estimar las curvas de los percentiles de peso es un problema de clasificación que involucra la estimación del percentil para cada momento temporal, calculando la varianza para cada punto gestacional. La estructura de efectos aleatorios nos permitió estimar la varianza y calcular los percentiles de peso en cada edad gestacional. Los resultados de este trabajo proporcionan una herramienta clínica validada que ofrece valores de diferentes puntos de corte (percentiles de peso). Este análisis permite al clínico identificar el comportamiento del crecimiento fetal, así como detectar fetos SGA o LGA considerando estos puntos de corte estimados, lo que puede ayudar a prevenir posibles efectos adversos.

A continuación se presenta el artículo publicado, derivado de este trabajo.

European Journal of Obstetrics & Gynecology and Reproductive Biology 253 (2020) 238–248

Contents lists available at ScienceDirect

European Journal of Obstetrics & Gynecology and Reproductive Biology

journal homepage: www.elsevier.com/locate/ejogrb

Full length article

A cohort study of fetal growth in twin pregnancies by chorionicity: comparison with European and American standards

Ricardo Savirón-Cornudella[a], Luis M. Esteban[b,*], Rocío Aznar-Gimeno[c], Faustino R. Pérez-López[d], Marta Chóliz Ezquerro[e], Peña Dieste Pérez[e], José M. Campillos Maza[e], Gerardo Sanz[f], Berta Castán Larraz[g], Mauricio Tajada-Duaso[e]

[a] Department of Obstetrics and Gynecology, Villalba General Hospital, Camino de Moralzarzal M-608 Km, Calle Alpedrete 41, 28400 Collado Villalba, Madrid, Spain
[b] Department of Applied mathematics, Escuela Universitaria Politécnica de La Almunia, Universidad de Zaragoza, Calle Mayor 5, 50100, La Almunia de Doña Godina, Zaragoza, Spain
[c] Deparment of BigData and Cognitive systems. Instituto Tecnológico de Aragon, ITAINNOVA, María de Luna 7-8, 50018, Zaragoza, Spain
[d] Department of Obstetrics and Gynecology, University of Zaragoza, Faculty of Medicine and Instituto de Investigación Sanitaria Aragón, Domingo Miral s/n, 50009, Zaragoza, Spain
[e] Department of Obstetrics and Gynecology, Miguel Servet University Hospital, Isabel La Católica 3, 50009, Zaragoza, Spain
[f] Department of Statistical Methods and Institute for Biocomputation and Physics of Complex Systems-BIFI, University of Zaragoza, Calle Pedro Cerbuna 12, 50009, Zaragoza, Spain
[g] Department of Obstetrics and Gynecology, San Pedro Hospital, Calle Piqueras 98, 26006, Logroño, La Rioja, Spain

ARTICLE INFO

Article history:
Received 1 June 2020
Received in revised form 3 August 2020
Accepted 21 August 2020

Keywords:
estimated fetal weight
estimated percentile weight
fetal growth
small for gestational age
twin gestations
chorionicity
ultrasound

ABSTRACT

Objective: To develop fetal growth standards for twin gestations by placental chorionicity in a Spanish population and compare them with European and American standards to estimate the suitability of their use in clinical practice.

Study design: This was a retrospective cohort study of 518 twin pregnancies, 435 dichorionic-diamniotic and 83 monochorionic-diamniotic, performed between January 2012 and December 2017. A total of 4,783 and 1,455 estimated fetal weights were considered from the 17th to the 37th week of gestation, using multilevel models, to build dichorionic-diamniotic and monochorionic-diamniotic standards, respectively. The percentages of small and large for gestational age were calculated as a model adjustment measure and adjustment to the studied data and the values provided by our model were compared against those of six European and American twin standards and three singleton standards. Correlation analyses between percentile predictions were performed using Cohen kappa coefficient. The predictive ability to detect small for gestational age was also provided by the sensitivity and positive predictive value.

Results: We found slight differences between standards by chorionicity, being dichorionic-diamniotic percentiles slightly higher than monochorionic-diamniotic ones from the 17th to 37th weeks' gestation. For dichorionic-diamniotic cases, both our standard (9.8-8.2) and that of Grantz (8.2-10.5) showed good adjustments for the 10th and 90th percentiles while the other compared standards underestimated or overestimated them. For monochorionic-diamniotic cases, both our standard (10.2-8.5) and that of Shivkumar (11.4-6.8) had the most suitable adjustment. The correlation analysis between small and large for gestational age cases provided by standards, showed clear differences among them. Kappa's coefficient showed a substantial agreement between both Ananth (0.7) and Stirrup (0.69) dichorionic-diamniotic cases and our standard. There was also a substantial agreement between the Shivkumar (0.77) standard and our results for monochorionic-diamniotic cases. The correlation was moderate for all other comparisons.

Conclusions: Our model showed a good adjustment to the studied population. There are clear differences among small and large for gestational age cases provided by twin standards in our studied population. The twin growth standards depend on the population characteristics and model structure. We found the use of singleton standards for twin pregnancies inadequate.

* Corresponding author at: Escuela Universitaria Politécnica de La Almunia, Calle Mayor 5, 50100, La Almunia de Doña Godina, Zaragoza, Spain
E-mail addresses: rsaviron@gmail.com (R. Savirón-Cornudella), lmeste@unizar.es (L.M. Esteban), raznar@itainnova.es (R. Aznar-Gimeno), faustino.perez@unizar.es (F.R. Pérez-López), martacholiz@gmail.com (M.C. Ezquerro), pdpe88@gmail.com (P.D. Pérez), jmcampillos@salud.aragon.es (J.M. C. Maza), gerardo.sanz@unizar.es (G. Sanz), bcastan@riojasalud.es (B.C. Larraz), mtajadad@gmail.com (M. Tajada-Duaso).

https://doi.org/10.1016/j.ejogrb.2020.08.044
0301-2115/

INTRODUCTION

The European Perinatal Report has shown that the median rate for multiple pregnancies was 16.7 per 1000 women in 2015 [1]. The rise in assisted reproductive technology is the main explanation for the growing rates of multiple pregnancies over the past decades. Multiple pregnancies carry higher risks of adverse perinatal outcomes (APOs), including substantially higher rates of gestational hypertensive disorders, preterm birth, and cesarean deliveries as compared to singleton gestations [2].

Different fetal growth standards calculated with the Hadlock formula [3] have been reported to determine the estimate percentile weight (EPW) in twins based on the birthweight and the estimated fetal weight (EFW) [4–10]. Twin standards differ from singleton ones from at least week 28 to 32 of gestation as reported for both birthweight and EFW standards [5,6,10,11]. When singleton standards are used, up to 30-40% of twin gestations are considered as small-for-gestational-age (SGA) [4,10,11]. Therefore, twin standards should be used in twin gestations, situation that would most likely reduce in these cases many unnecessary interventions [5,12] and preterm births [2]. These would also be more effective at identifying twin pregnancies at risk of intrauterine fetal death [11] and neonatal morbidity and mortality [13].

Differentiation by chorionicity has been proposed in twin standards since birthweight differs in dichorionic-diamniotic (DC) and monochorionic-diamniotic (MCDA) twin pregnancies, being lower in most studies for MCDA [4,6,7,10,14]. Perinatal morbidity and mortality also differ in the two groups, being higher in MCDA due to an increased risk of preterm birth and growth discordance [15] and, also associated with their specific pathologies (twin-to-twin transfusion syndrome and twin anemia polycythemia syndrome).

The aim of this study was to develop a fetal growth standard for twin gestations by placental chorionicity in a Spanish population and to compare the performance of European and American twin standards through estimates of SGA and LGA cases in our cohort.

METHODS

This was a retrospective study that analyzed data of twins delivered between January 2012 and December 2017. The original database included 578 mothers and 1156 infants. The inclusion criteria were as follows: (i) live pregnancies that received prenatal care since the first trimester of gestation at the Miguel Servet University Hospital (MSUH), in the Northeast part of Spain, (ii) fetal ultrasound assessment between **17** and 37 weeks of gestational age; and (iii) deliveries of live fetuses above the 24th gestational week. Cohort members were excluded if they showed karyotype abnormalities, twin-twin transfusion syndrome (TTTS), major congenital malformation or incomplete data. Because of their rarity, 5 monochorionic-monoamniotic cases were excluded from our analysis. Moreover, pregnancies with less than 4 ultrasound measurements by fetus were not considered as they can provide non robust estimations of the parameters of the model. All data were checked and a total of 316 errors in measurement or biologically implausible values were also excluded. Analyses were finally performed for 518 mothers and 1036 infants (Fig. 1).

Ultrasound screening in the gestations were performed at the Ultrasound and Prenatal Diagnosis Unit of the mentioned hospital, using Voluson 730 Expert, E6 or E8 model ultrasound machines (General Electric, Healthcare, Zipf, Austria). Gestational age in days was estimated by adjusting the last menstrual period (LMP) based on measurement of the crown–rump length (CRL) by the first trimester ultrasound [16], at which time the determination of chorionicity was also performed to split the database into DC and MCDA cases.

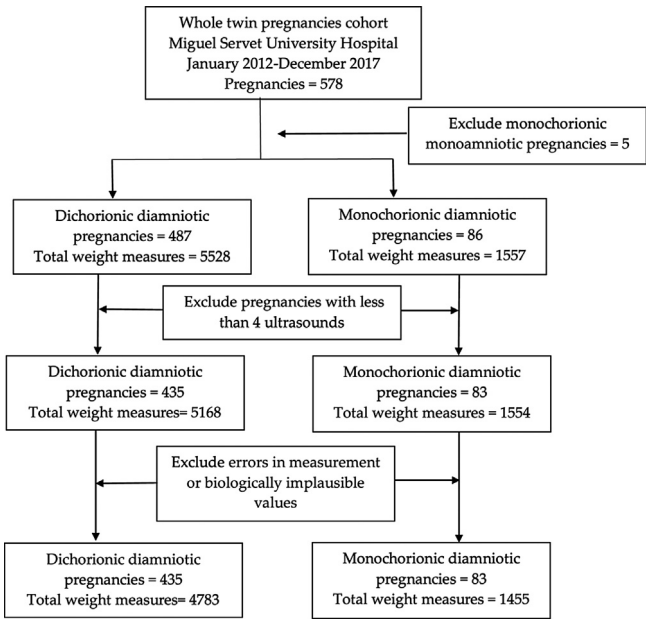


Fig. 1. Study participants selection sample.

EFW was calculated with the Hadlock's formula (biparietal diameter, head circumference, abdominal circumference and femur length) [3] to build our twin growth standards. The rest of the standards with which we made the comparison also use Hadlock's formula to estimate fetal weights, with the exception of INTERGROWTH-21st that uses the Stirnemann formula [17]. Hadlock's formula was also used to estimate weight percentiles of the Stirrup et al. [11] standard, given estimated biometric measurements.

Statistical analysis

Data were extracted and the maternal-fetal characteristics and the weight discordance intra-fetuses were descriptively analysed. Continuous variables were reported as medians with interquartile ranges whereas qualitative variables were expressed as frequencies and percentages. Differences between chorionic groups were calculated using the Mann-Whitney test or the chi-squared test, as appropriate. Our data set consisted of repeated measures of EFW for each and both fetuses of the same mother. This intrinsic hierarchy of data violates the statistical assumption of independence. The use of multilevel models allowed us to solve this non-independence of the data in order to model the relationship between gestational age and the estimated weight of these correlated observations, considering them as random effects [18,19].

The distribution of fetal weight is known to be nonlinear across gestational age, with a higher growth rate in the second trimester and a lower one in the third trimester [20]. To account for this nonlinearity, we modeled the gestational age as a fixed effect using restricted cubic splines with 5 knots [21]. EFW measurements were log transformed to ensure the normality distribution assumption by week. The Akaike criterion was used for the variable selection process and the models were compared with analysis of variance test. The R squared statistic [22] was used to analyze the goodness of fit of the model.

To provide EPW, the variance was estimated for every gestational week from the random effects of the multilevel models, allowing calculation of percentiles 5, 10, 50, 90 and 95. Besides, to estimate the performance of twin growth standards, we calculated the percentage of SGA, defined as EPW under 10%, and LGA, defined as EPW over 90%, from the DC and MCDA standards in our cohort. Comparisons were performed with the non-customized twin standards of Ananth et al. [4], Araujo et al. [6], Gabbay-Benziv et al. [9], Grantz et al. [8],

Shivkumar et al. [7] and Stirrup et al [10] complemented by Savirón-Cornudella et al. [23], INTERGROWTH-21st [19] and WHO singleton [24] standards. Moreover, to analyze a period closer to delivery, the study was also carried out considering only the last weeks of pregnancy from week 32. In addition, a correlation analyses between the SGA, AGA and LGA cases provided by twin and singleton standards and our standard was performed using the Cohen kappa coefficient. Finally, taking as a positive class the SGA classified by our model during the gestational period, sensitivity and positive predictive value (PPV) was also provided as a measure of the ability to predict SGA, which is the most vulnerable population to have problems. Specifically, the tp (true positives) are the fetuses classified as SGA by both our model and the model with which we compare and the fp (false positives) are the fetuses classified as SGA by the model with which we compare but they were not classified as SGA by our model.

Analyses were performed using the R v.3.5.2 programming language and the lme4 package (The R Foundation for statistical computing, Vienna, Austria) [25]. For the use of the model we developed the R package **PTwins** freely available in CRAN repository https://CRAN.R-project.org/package=PTwins and we provide an app for clinical use https://ptwins.shinyapps.io/PTwins/.

Ethical approval

The study was approved on November 21, 2018, by the Clinical Research Ethics Committee of Aragon (CEICA, PI 2018/333).

RESULTS

Of the 518 pregnancies included in the study, 435 were DC cases with a median of 6 (interquartile range 5-6) weight measures by fetus and 83 were MCDA cases with a more detailed follow-up shown by a median of 9 (interquartile range 7-11) weight measurements. Table 1 displays maternal and fetal characteristics. For MCDA deliveries, 41 (49.4%) were preterm (32 th-37th week) and 4 (4.8%) were very preterm (28 th-32th week). In the DC case, 176 (40.5%) deliveries were preterm and 2 (0.5%) were very preterm. The percentage of nulliparous women was significantly different for DC and MCDA groups, being 78% and 67%, respectively. The fetuses of nulliparous women were associated with significantly lower birthweights than those of non-nulliparous women, both in DC and in MCDA pregnancies ($P < 0.001$ and $P = 0.015$, respectively), with a more noticeable difference observed in the DC cases. The most remarkable difference between twin pregnancies corresponded to the proportion of *in vitro* fertilizations, which was above 65% for DC and below 25% for MCDA pregnancies, respectively. However, the difference in birthweight between pregnancies with and without *in vitro* fertilization was not significant either in DC or MCDA cases ($P = 0.74$ and $P = 0.49$, respectively). Median maternal weight and age were similar, rounding 63 kg, and 34 years, respectively. There was a significant increase of 4 days in the gestational age at birth for the DC group, corresponding to a non-significant difference of 100 g in birthweight. There were no differences neither in sex ratios or weight discordance ratios of fetuses between DC and MCDA groups.

Tables 2 and 3 display comparisons of maternal-fetal characteristics (parity, body mass index, maternal age, birth weight, gestational age at delivery and weight percentile) of twin standards included in this study. In addition, we report the weight formula and the models used in each growth standard.

To build the standards, 4783 observations were finally considered for the DC cases and 1455 for the MCDA cases. The final growth model included two random effects, the first one, a random slope between pregnancy and a cubic polynomial structure of gestational age, and the second one, a random intercept for each fetus. The fixed and random effects coefficients are shown in Tables 4 and 5. The marginal R squared and conditional R squared [22] were 0.9740616 and 0.9942701 respectively for DC standard and 0.9807571 and 0.9962406 respectively for MCDA standard. Fig. 2 shows the 3rd, 10th, 50th, 90th and 97th percentile curves for the DC and MCDA pregnancies, where a reasonable adjustment of our twin standards to the data is observed. Although we found slight differences between the 50th percentile twin growth standards by chorionicity, with respect to the 10th and 90th percentiles (P10, P90, respectively) the differences are more significant, being the P10-P90 interval wider for DC cases.

Tables 6a and 6b shows the number of ultrasound measures (N) at every gestational age, from the 17th to the 37th week, as well as the mean and standard deviation (SD) of the ultrasound estimated weights of our population. The 3rd, 5th, 10th, 50th, 90th, 95th and 97th percentiles estimated for DC and MCDA growth standards are also shown at every gestational age in Tables 6a and 6b. In DC twins, there was an increase in ultrasounds performed in the last weeks of

Table 1
Maternal and fetal characteristics of dichorionic and monochorionic-diamniotic pregnancies in studied Spanish cohort.

Pregnancy characteristics	Dichorionic pregnancies (n_{dc} = 435)[†]	Monochorionic diamniotic pregnancies (n_{mc} = 83)[‡]	p-value
Maternal age (years)	34.1 (30.4-36.6)	34.2 (30.2-36.4)	.968
Parity (nulliparous)	340 (78.3%)	55 (67.1%)	.028
In vitro fertilization (yes)	289 (66.4%)	19 (22.9%)	< .001
Smoking habit (yes)	50 (11.5%)	16 (19.3%)	.077
Maternal weight at first trimester (kg)	62 (57-71)	64 (58-72)	.266
Maternal height (cm) (n_{dc} = 295, n_{mc} = 58)	165 (160-168)	163 (160-167)	.327
Paternal height (cm) (n_{dc} = 289, n_{mc} = 59)	178 (172-183)	175 (170-179)	.003
Gestational age at birth (days)	261 (252-266)	257 (245-260)	< .001
Birthweight (grams)	2470 (2170-2740)	2360 (2010-2570)	.266
Weight discordance [§]:			
≥5%	1622 (65.75%)	435 (60.42%)	0.009691
≥10%	826 (33.48%)	256 (35.56%)	0.3226
≥15%	407 (16.5%)	139 (19.3%)	0.08856
≥20%	186 (7.53%)	65 (9.03%)	0.2203
≥25%	87 (3.53%)	30 (4.17%)	0.4896
≥30%	33 (1.34%)	9 (1.25%)	1
Sex (female)	456 (52.4%)	94 (56.6%)	.362

[†] n_{dc}: number of dichorionic pregnancies,
[‡] n_{mc}: number of monochorionic pregnancies.
[§] Formula weight discordance: 100 × (weight of larger twin − weight of smaller twin)/weight of larger twin. Calculated for each ultrasound measurement

Table 2
Dichorionic twin standards characteristics.

	Our cohort*	Ananth et al†	Araujo et al†	Gabbay-Benziv et al*,§	Grantz et al†	Shivkumar et al†	Stirrup et al**
Number of twin gestations	435	1030	176	1427	171	540	1802
Nulliparous (%)	78.3	45‡	unknown	unknown	56.1	51.40	unknown
Body mass index (kg/m^2)	23 [21,26]	unknown	unknown	29.8 (15.7-72.9)	28.6 ± 7.0	24.2 ± 4.8	unknown
Maternal age (years)	34.1 [30.4,36.6]	unknown	28.75 ± 6.86	33 (17-50)	31.6 ± 6.1	33 ± 5.1	unknown
Birthweight (grams)	2470 [2170,2740]	2356 ± 652	unknown	unknown	2376 [1960, 2879] ¶	2669	unknown
Gestational age at delivery (weeks)	37.3 [36, 38]	35.3 ± 3.4	unknown	unknown	35	37 ± 1	unknown
Weight percentile at 36 th week (g) (Percentiles 50th [10 th -90 th])	2408 [2034, 2852]	2456 [1985, 2928]	2457 [1969,2946]	2308 [1753, 2863]	2622 [2106,3138]	2691 [2271, 3190]	2596[2082, 3202]
Weight's formula	Hadlock 1985	Birthweight	Hadlock 1985	Hadlock 1985	Hadlock 1985	Hadlock 1985	Hadlock 1985
Model	Linear Mixed model	Restricted spline smoothing	Polynomial regression	Linear Mixed model	Linear Mixed model	Linear Mixed model	Linear Mixed models
Population (City, Country)	Zaragoza, Spain	Brighton, UK	São Paulo, Brazil	Baltimore, US	8 US sites	Montreal, Canada	Southwest Thames region, UK
Data range	2012-2017	1990-1996	unknown	2006-2016	2012-2013	1996-2006	2001-2010

Only first author given for each study.
* Median [Interquartile range].
† Mean ± SD. ‡Characteristic for all twins.
§ Characteristics for all twins except N, birthweight and weight percentile at 36th week.
¶ Median [Interquartile range] at 35th week.
** Twin standard for biometric variables. We used Hadlock 1985 formula to calculate the weight percentiles for each gestational age.

Table 3
Monochorionic-diamniotic twin standards characteristics in our cohort, and studies performed in other populations.

	Our cohort*	Ananth et al†	Araujo et al†	Gabbay-Benziv et al*,§	Shivkumar et al†,§	Stirrup et al¶
Number of twin gestations	83	272	157	688	102	300
Nulliparous (%)	67.1	45.0‡	unknown	unknown	51.4	unkown
Body mass index (kg/m^2)	23.5 [21.3, 26.0]	unknown	unknown	29.8 (15.7-72.9)	24.2 ± 4.8	unkown
Maternal age (years)	34.2 [30.2, 36.4]	unknown	29.3 ± 6.8	33 (17-50)	33 ± 5.1	unkown
Birthweight (g)	2360 [2010, 2570]	2139 ± 672	unknown	unknown	2471	X
Gestational age at delivery (weeks)	36.7 [35, 37.1]	34.3 ± 3.9	unknown	unknown	37 ± 1	unkown
Weight percentile at 36 th week (g) (Percentiles 50th [10 th -90 th])	2391 [2040, 2804]	2362 [1887, 2837]	2394 [1833, 2956]	2189 [1666, 2713]	2561 [2123, 3089]	2530[1975, 3189]
Weight's formula	Hadlock 1985	Birthweigt	Hadlock 1985	Hadlock 1985	Hadlock 1985	Hadlock 1985
Model	Linear Mixed model	Restricted spline smoothing	Polynomial regression	Linear Mixed model	Linear Mixed model	Linear Mixed models
Population (City, Country)	Zaragoza, Spain	Brighton, UK	São Paulo, Brazil	Baltimore, US	Montreal, Canada	Southwest Thames region, UK
Data range	2012-2017	1990-1996	unknown	2006-2016	1996-2006	2001-2010

Only first author given for each study.
* Median [Interquartile range].
† Mean ± SD.
‡ Characteristic for all twins.
§ Characteristics for all twins except N, birthweight and weight percentile at 36th week.
¶ Twin standard for biometric variables. We used Hadlock 1985 formula to calculate the weight percentiles for each gestational age.

pregnancy. By contrast, in MCDA cases, the distribution of ultrasounds was more homogeneous over time, indicating a more detailed follow-up. There was evidence of greater values in EPW for the DC twins until the 37th week of gestation.

We compared our reference with those previously reported by Ananth et al [4], Araujo et al [6], Gabbay-Benziv et al [9], Grantz et al [8], Shivkumar et al [10] and Stirrup et al [10]. Figs. 3 and 4 compare the growth curves for twin standards and show differences along with the gestational age, presenting varied growth rates. Figs. 5 and 6 display the standard growth curves for singletons.

Adjustment analysis along the gestational period (from 17th and 37th week) is shown in Fig. 7. For DC cases, Grantz et al [8] standard and our standard showed a good adjustment, with estimated percentages of SGA and LGA of 8.2-10.5 and 9.8-8.2, respectively. Araujo et al [6] and Gabbay-Benziv et al [9] standards underestimated 10th and 90th percentiles and the Ananth et al [4] and Stirrup et al [10] standards underestimated 10th percentile and overestimated the 90th percentile. The Shivkumar et al [7], and the singleton standards, overestimated the 10th and 90th percentiles. For the MCDA cases, our standard had a reasonable adjustment, with estimated percentages of SGA and LGA of 10.2 and 8.5, respectively. The Shivkumar et al [7] standard had the most suitable adjustment, considering around 11% of our population below its 10th percentile. The Ananth et al [4], Araujo et al [6], Gabbay-Benziv et al [9] and Stirrup et al [10] standards underestimated the 10th and 90th percentiles and the singleton standards overestimated 10th and 90th percentiles. The Ananth et al [4] and

Table 4
Fixed and random effects coefficients for dichorionic twins.

Dichorionic twins	Parameter	Coefficient	p-value
	Fixed-effects		
	Intercept	7.424321	< .0001
	First gestational age spline basis term	1.109570	< .0001
	Second gestational age spline basis term	−0.899378	< .0001
	Third gestational age spline basis term	1.392611	< .0001
	Fourth gestational age spline basis term	−0.998825	< .0001
	Random-effects		
Fetus level			< .0001
	Variance, intercept	.0055716999	
Pregnancy level			< .0001
	Variance, intercept	.0051698540	
	Variance, first degree gestational age polynomial term	.0024051903	
	Variance, second degree gestational age polynomial term	.0004668487	
	Variance, third degree gestational age polynomial term	.0003763530	
	Covariance, first degree gestational age polynomial term, Intercept	.0010606255	
	Covariance, second degree gestational age polynomial term, Intercept	−0.0006350283	
	Covariance, third degree gestational age polynomial term, Intercept	−0.0002599014	
	Covariance, second degree gestational age polynomial term, first degree gestational age polynomial term	.0001133536	
	Covariance, third degree gestational age polynomial term, first degree gestational age polynomial term	−0.0006754863	
	Covariance, third degree gestational age polynomial term, second degree gestational age polynomial term	.0001165476	
	Residual variance	.0033025276	

Table 5
Fixed and random effects coefficients for monochorionic diamniotic twins.

Monochorionic twins	Parameter	Coefficient	**p-value**
	Fixed-effects		
	Intercept	7.21712	< .0001
	First gestational age spline basis term	1.32495	< .0001
	Second gestational age spline basis term	−0.87209	< .0001
	Third gestational age spline basis term	1.34849	< .0001
	Fourth gestational age spline basis term	−0.97392	.0003
	Random-effects		
Fetus level			<.0001
	Variance, intercept	.005955849	
Pregnancy level			< .0001
	Variance, intercept	.005535426	
	Variance, first degree gestational age polynomial term	.002235781	
	Variance, second degree gestational age polynomial term	.0004520071	
	Variance, third degree gestational age polynomial term	.0002315112	
	Covariance, first degree gestational age polynomial term, Intercept	−0.0009080299	
	Covariance, second degree gestational age polynomial term, Intercept	−0.0007691504	
	Covariance, third degree gestational age polynomial term, Intercept	.0001993174	
	Covariance, second degree gestational age polynomial term, first degree gestational age polynomial term	.0001268664	
	Covariance, third degree gestational age polynomial term, first degree gestational age polynomial term	−0.0004766251	
	Covariance, third degree gestational age polynomial term, second degree gestational age polynomial term	.0000267061	
	Residual variance	.002948896	

Stirrup et al [10] standards had a good adjustment to estimate the 90th percentiles, with percentages of LGA of 10 and 10.3, respectively.

Table 7 show the adjustment analysis from 32th week of gestation. For DC cases, while our standard showed a good adjustment, for this period, Ananth et al [4] and Stirrup et al [10] standards are the best adjustment showed for the estimation of 10^{th} percentiles. For MCDA cases, our standard and Shivkumar et al [7] kept had the most suitable adjustment.

The correlation analysis between SGA and LGA cases provided by standards is shown in Table 8. Both the sensitivity and positive predictive values (PPV) were over 60% in the Ananth et al [4], Araujo et al [6], Grantz et al [8], Shivkumar et al [7] and Stirrup et al [10] standards for DC cases and over 80% only in the Shivkumar et al [7] standard for MCDA cases. In particular, 92.6% of the fetuses that were classified as SGA by our model were also classified as SGA by the Shivkumar et al [7] standard (sensitivity), and 80.6% of the fetuses that were classified as SGA by the standard were also classified as SGA by our model (PPV). Kappa's coefficient showed substantial correlation with Grantz et al [8], Stirrup et al [10]] and Ananth et al [4] standards and our standard in DC cases and between Shivkumar et al [7] and our standard for MCDA cases. The remaining comparisons presented a moderate agreement. Regarding the singleton standards, the Cohen-Kappa coefficients showed a moderate correlation in all cases. Therefore, it can be concluded that the correlation between SGA and LGA cases provided by standards varied in our study.

DISCUSSION

Although a variety of twin standards by chorionicity have been reported in European and American populations, there is no other work with such an extensive comparative analysis. Our study includes the generation and validation of non-customized standards that allow us to construct growth curves and reference tables of EFW for the twin population according to their placenta. The present study also offers the comparison and validation of various state of the art standards with regard to our population. We have used multilevel models to build fetal weight standard by chorionicity that consider both, the relation between the average

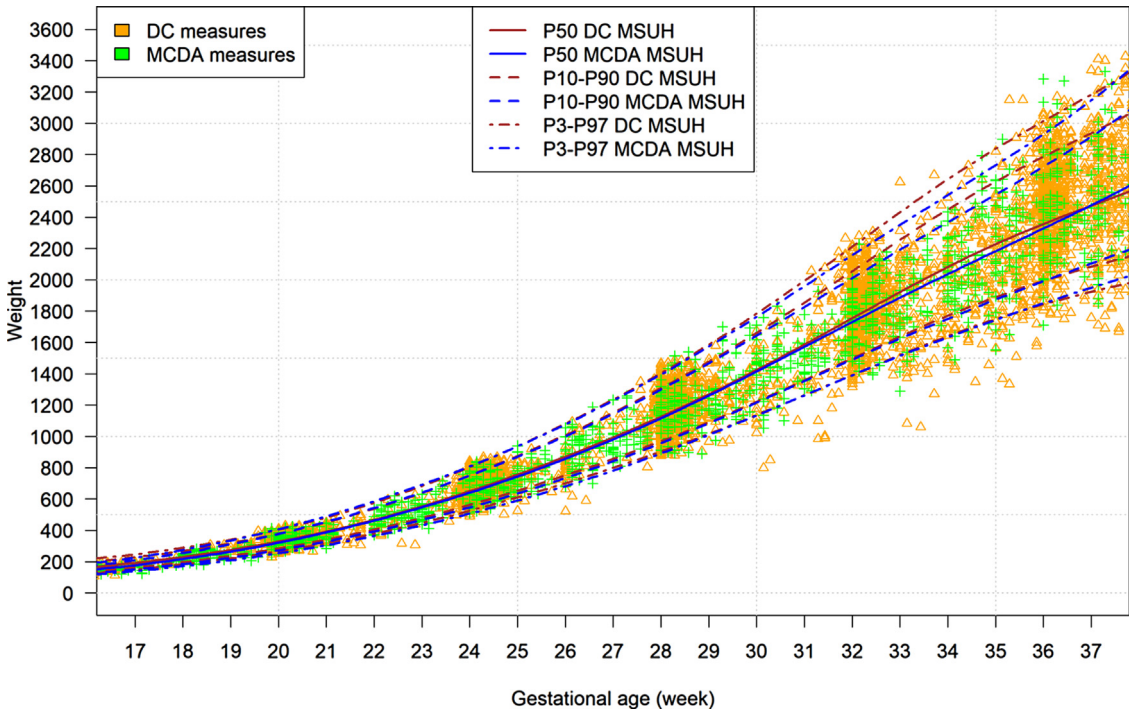


Fig. 2. Miguel Servet University Hospital (MSUH) standard charts for dichorionic (DC) and monochorionic-diamniotic (MCDA) twin pregnancies for percentile 50 (P50) and percentile 10-90 (P10-P90).

Table 6a
Percentiles distribution in dichorionic pregnancies

Gestational age	Dichorionic-diamniotic pregnancies									
(weeks)	†n_{mc}	Mean	SD	3rd	5th	10th	50th	90th	95th	97th
17	89	192.8	13.1	150	155	162	192	228	240	247
18	8	267.8	36.4	183	189	197	230	268	281	289
19	79	317.2	34.9	222	228	237	275	318	331	341
20	701	340.7	33	266	273	285	328	379	394	405
21	49	397.3	46.6	318	327	340	392	452	470	483
22	25	461.8	78.6	379	389	405	467	539	561	576
23	65	624.9	63.8	448	460	479	553	639	666	683
24	708	677.2	71.2	525	539	562	650	752	783	805
25	39	739.1	81.2	610	626	653	756	875	912	937
26	53	852.5	131.3	701	720	751	870	1009	1052	1081
27	61	1110.1	120.1	799	821	856	993	1153	1202	1236
28	680	1171.8	124	904	929	969	1126	1308	1365	1403
29	74	1290.5	130.9	1016	1045	1091	1270	1477	1542	1586
30	53	1375	226.6	1137	1169	1221	1424	1660	1734	1784
31	87	1675.7	284.2	1263	1300	1358	1587	1855	1938	1995
32	662	1809.5	188.1	1392	1433	1499	1755	2055	2149	2213
33	124	1861.8	292.7	1520	1565	1638	1922	2256	2361	2432
34	131	2028	291.9	1641	1691	1770	2083	2450	2565	2643
35	223	2242.4	344	1750	1804	1890	2228	2627	2753	2838
36	582	2483.4	275	1843	1901	1993	2357	2787	2923	3014
37	290	2583.3	360.3	1923	1985	2084	2474	2938	3084	3183

† n_{mc}: number of dichorionic pregnancies

attained fetal weight and GA, and how variability (standard deviation) changes with gestational age, therefore EFW is more adjusted to gestational age. We decided to omit from our analysis some studies that analyzed fetal sex and other maternal characteristics such as predictive factors of birthweight in twin standards [26,27], since recent publications show that customization does not provide a greater capacity to detect APOs using SGA as a predictive factor in singleton standards [28,29], and makes a more difficult comparison for the variety of variables included.

Focusing on MCDA pregnancies, Shivkumar et al [7] shown a reasonable adjustment for our population, especially to predict SGAs with only a small overestimation of LGA. In DC cases, percentile predictions showing a large proportion of SGA and a small proportion of LGA cases. The Shivkumar et al [7] cohort had less nulliparous pregnancies than the rest of standards, which could be relevant for DC twins. In addition, the multilevel structure includes random slopes for each pregnancy and fetus level. By contrast, in the MCDA standard the mixed model only includes a random slope at the fetus level, and there is good agreement with our standard with a similar structure.

The Grantz et al [8] standard only developed for DC cases, showed a reasonable adjustment although with a low correlation in SGA and LGA cases with our standard. However, there were clear differences in BMI and the percentage of nulliparous women

Table 6b
Percentiles distribution in monochorionic-diamniotic pregnancies.

Gestational age		Monochorionic-diamniotic pregnancies								
(weeks)	† n_{mc}	Mean	SD	3rd	5th	10th	50th	90th	95th	97th
17	119	173.3	27.4	139	144	151	177	209	219	225
18	93	233.1	25.5	171	176	184	217	255	267	275
19	30	289.6	54.4	208	215	225	264	311	326	335
20	117	345.8	36.3	253	261	273	321	377	395	407
21	20	377.7	53	305	314	329	387	455	476	490
22	84	487.3	56.3	365	376	393	462	543	568	585
23	44	559.4	85	432	445	466	547	642	672	692
24	98	680.2	479.3	507	522	546	640	751	785	809
25	47	758.1	84	590	608	636	744	870	910	937
26	70	912.9	106.1	682	702	733	857	1001	1046	1077
27	43	1000.1	111.5	782	805	841	981	1144	1195	1230
28	101	1181.9	151.3	892	917	958	1116	1300	1357	1396
29	51	1312.4	177.4	1009	1038	1084	1261	1467	1531	1575
30	61	1461.9	168.4	1134	1166	1216	1414	1644	1716	1764
31	48	1567.2	146.3	1262	1297	1354	1573	1827	1906	1960
32	91	1788	185.9	1390	1429	1491	1732	2011	2099	2157
33	40	1865.7	238.2	1515	1557	1624	1887	2192	2287	2351
34	87	2090.2	255.5	1633	1679	1752	2037	2369	2472	2542
35	60	2265	288.9	1745	1794	1874	2183	2544	2656	2732
36	87	2454.4	315.4	1850	1904	1991	2329	2723	2847	2930
37	64	2510.2	329.6	1950	2009	2104	2477	2916	3053	3146

† n_{mc}: number of monochorionic pregnancies.

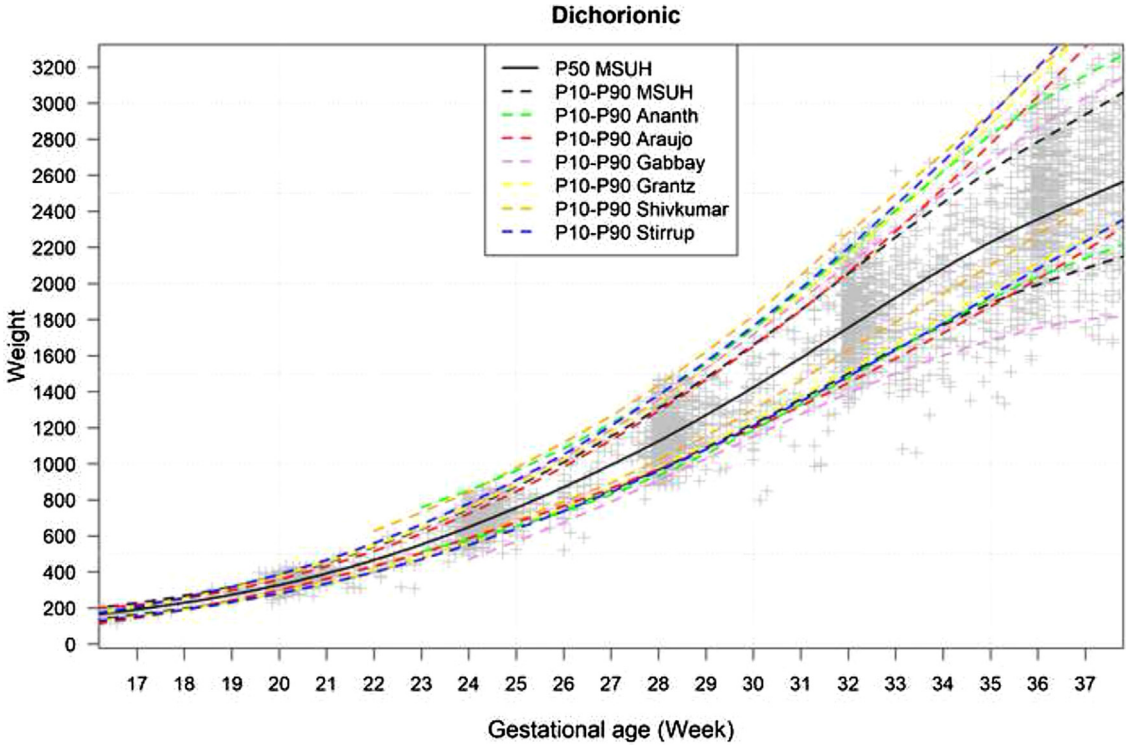


Fig. 3. Comparison between twin standards for dichorionic cases: Miguel Servet University Hospital (MSUH) percentile 50 (P50) and percentiles 10-90 (P10-P90) for all the standards.

cohort. We found substantial correlation with our standard and the Grantz standard built using mixed model structure that only includes random slope for the twin pair.

By contrast, the Gabbay-Benziv et al [9] and Araujo et al [6] standards clearly underestimate SGA and overestimate LGA cases in DC and MCDA models. The lack of information in the generation of the cohorts limits the comparison, but maternal age is clearly lower in the Araujo et al [6] study and BMI is greater in the Gabbay-Benziv et al [9] cohort.

Ananth et al [4] and Stirrup et al [10] underestimates both SGA and LGA cases in DC models, being in the Ananth study the percentage of nulliparous pregnancies and birthweight was lower than our cohort and unknown in Stirrup study. For both models, the 10th percentile shows a similar growth curve to ours with only slight differences, but there is a clear overestimation of the 90th percentile. For the Ananth study, we suspect that this gap is explained by the use of birthweights instead of EFW in this study. For the MCDA models, the underestimation of SGA cases is more

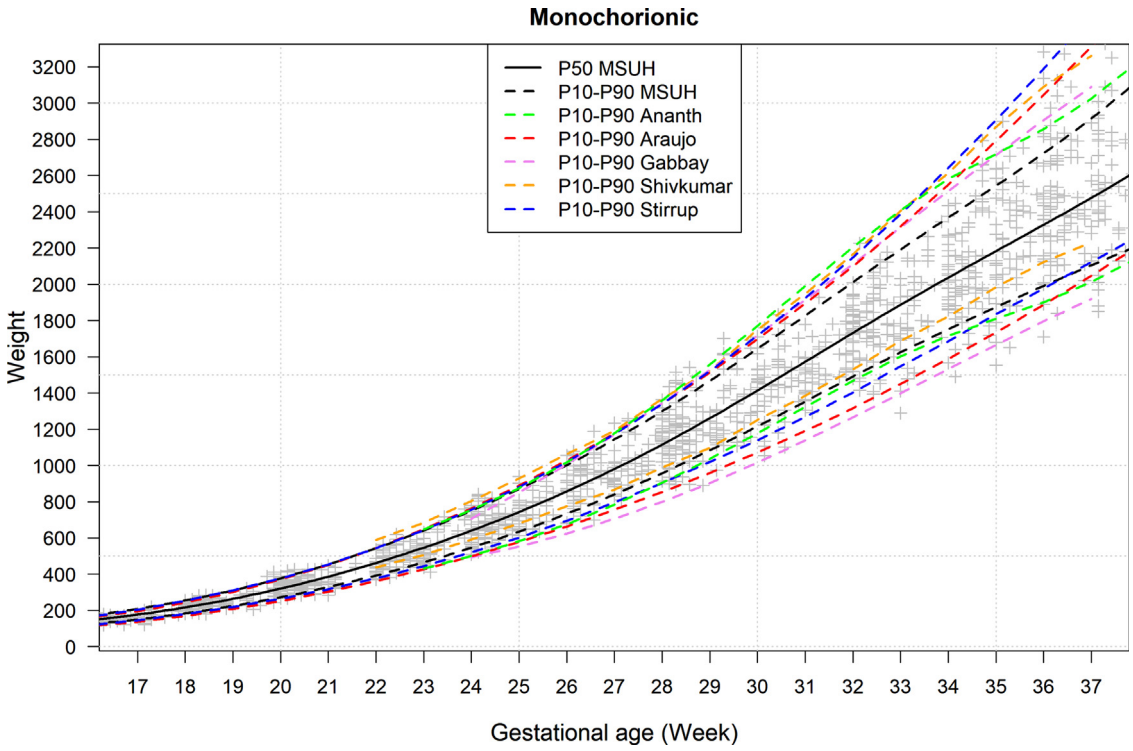


Fig. 4. Comparison between twin standards for monochorionic-diamniotic cases: Miguel Servet University Hospital (MSUH) percentile 50 (P50) and percentiles 10-90 (P10-P90) for all the standards.

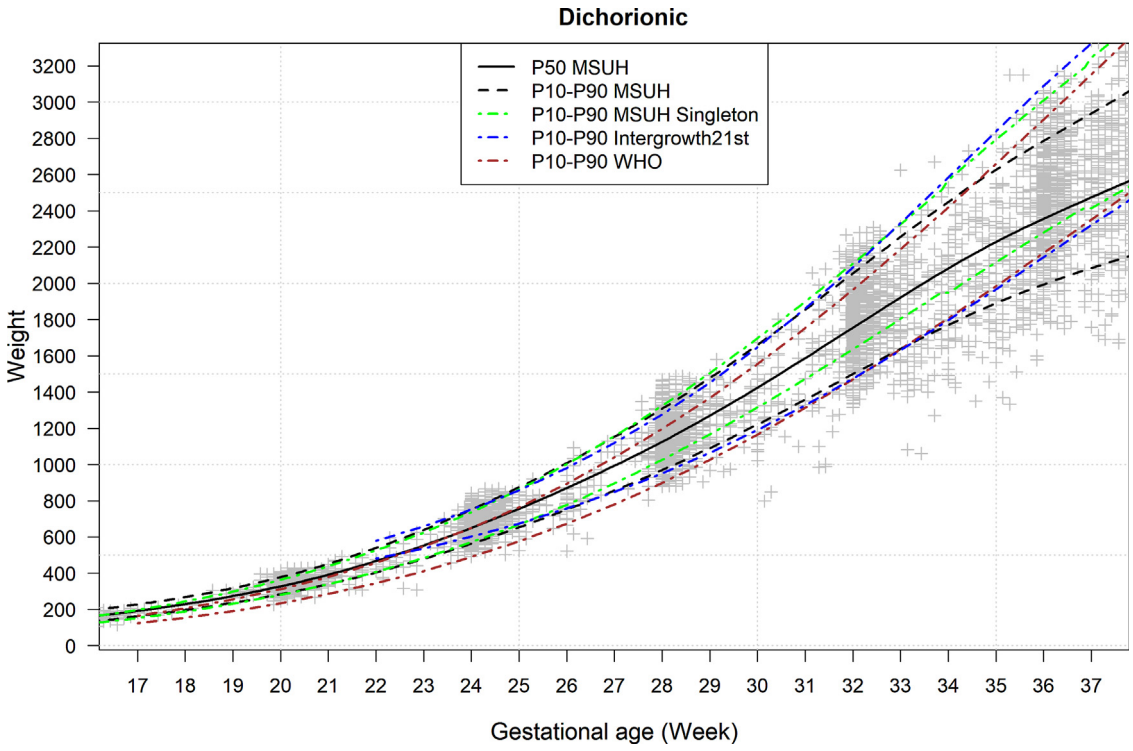


Fig. 5. Comparison between twin and singleton standards for the dichorionic cases: Miguel Servet University Hospital (MSUH) percentile 50 (P50) and percentiles 10-90 (P10-P90) for all the standards.

pronounced in both models, but with a very good adjustment for LGA cases.

Focusing in the growth curves showed in Figs. 3 and 4, it is remarkable that all the growth standards show significant differences in the last few weeks of gestational age in the DC cases; the slope of the curve increased for Araujo et al [6] and for Stirrup et al [10] and decreased for Gabbay-Benziv et al [9]. Also, in the adjustment study for the period 32^{th}-37^{th} week of gestional,

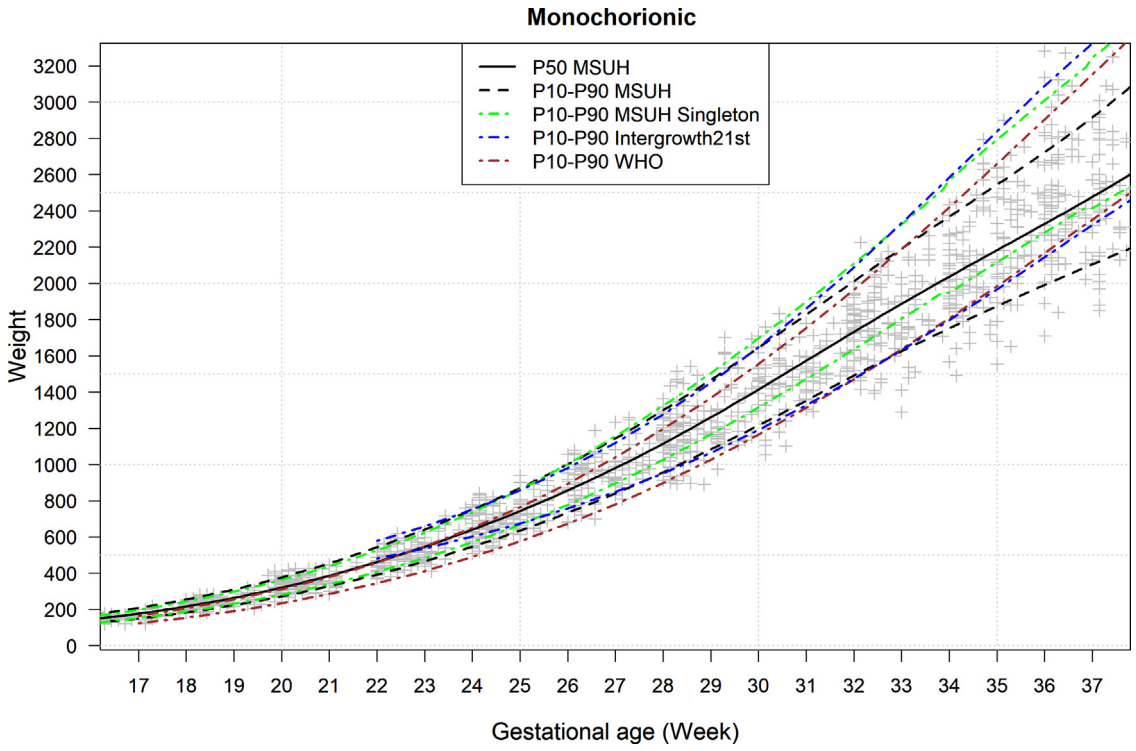


Fig. 6. Comparison between twin and singleton standards for the monochorionic-diamniotic cases: Miguel Servet University Hospital (MSUH) percentile 50 (P50) and percentiles 10-90 (P10-P90) for all the standards.

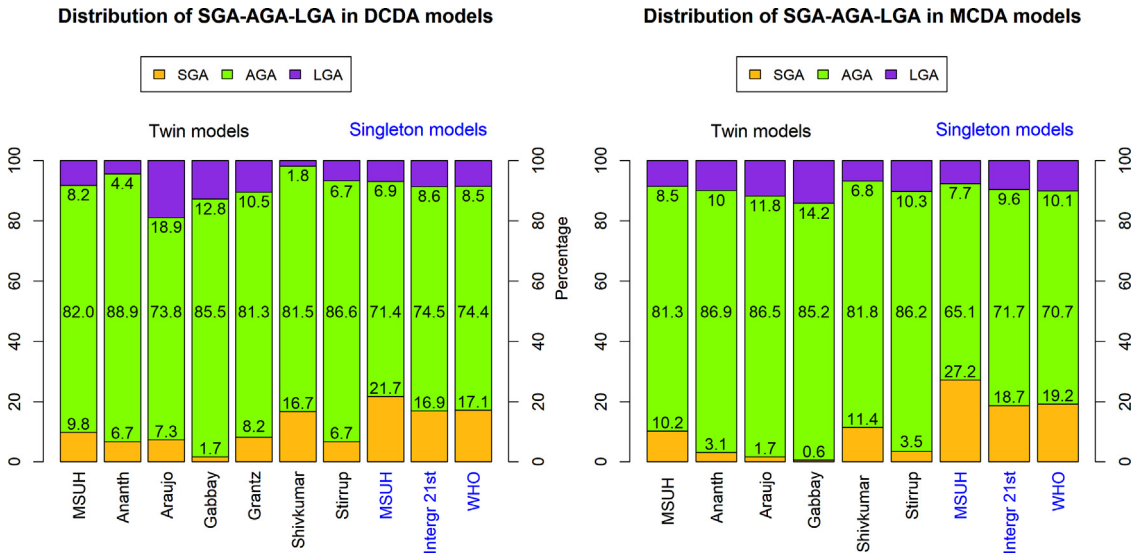


Fig. 7. Adjustment of twin and singleton standards for the Miguel Servet Univesity Hospital (MSUH) studied population. AGA: adequate for gestational age. DCDA: dichorionic-diamniotic. LGA: large for gestational age. MCDA: monochorionic-diamniotic. SGA: small for gestational age. WHO: World Health Organization.

age, there is an increased lack of adjustment for all models with the exception of our model, it means models failed when we approach to the delivery.

Regarding singleton gestations, differentiation with twins begins at 33-34 weeks of gestational age in DC pregnancies and at 33 weeks of gestational age in DC. In our comparison with singleton standards, WHO and INTERGROWTH-21st growth standards showed a similar performance, with a large proportion of SGA cases, but especially over the 32nd week of gestational age. The 90th percentile showed dissimilarity over the 30th week. This behavior was even more pronounced in the Saviron-Cornudella et al [23] singleton standard.

We found significant differences among small and large for gestational age cases provided by twin standards in our studied Spanish population. Twin standards showed differences depending on the population characteristics and by the model structure. Twins have to be stratified by chorionicity since dichorionic twins are heavier than monochorionic and the use of singleton standards for twin pregnancies results in an overestimation of SGA. On the other hand, while Hadlock's formula is the method that produces the smallest random errors [30], it is known that ultrasound fetal weights estimation has a measure of error commonly overestimating real weight, being less accurate in singleton than in twin pregnancies [31]. This limitation should be

Table 7
Distribution of SGA-LGA in twin standards from 32th week of gestational age.

	Dichorionic		Monochorionic-Diamniotic	
Twin Standards	SGA	LGA	SGA	LGA
Our standard	11,4	8,2	9,6	9,3
Ananth et al	9,31	4,1	5,2	5,2
Araujo et al	8,1	6,8	2,1	3,9
Gabbay-Benziv et al	2,1	8,9	1,1	5,6
Grantz et al	12,4	3,1	–	–
Shivkumar et al	21,7	0,68	10,7	1,9
Stirrup et al	10,88	1,64	6,2	2,6

Table 8
Correlation analyses between our cohort results with other twin standards, dichorionic and monochorionic-diamniotic, and with singleton standards.

		Dichorionic twin pregnancies					Monochorionic-diamniotic pregnancies				
		Our cohort					Our cohort				
Twin Standards		**SGA**	**AGA**	**LGA**	**Correlation**		**SGA**	**AGA**	**LGA**	**Correlation**	
Ananth et al*	**SGA**	256	5	0	K	0.70 (0.67-0.73)	31	0	0	K	0.54 (0.47-0.60)
	AGA	150	3177	154	Sens	63.1	68	769	29	Sens	36.5
	LGA	0	2	170	PPV	98.1	0	38	67	PPV	100
Araujo et al†	**SGA**	310	38	0	K	0.50 (0.48-0.53)	23	0	0	K	0.43 (0.37-0.49)
	AGA	162	3292	85	Sens	65.7	117	1037	30	Sens	16.2
	LGA	0	599	307	PPV	89.1	0	75	86	PPV	91.3
Gabbay-Benziv et al‡	**SGA**	67	0	0	K	0.54 (0.51-0.58)	6	0	0	K	0.44 (0.37-0.51)
	AGA	333	2953	4	Sens	16.8	86	715	15	Sens	6.5
	LGA	0	176	316	PPV	100	0	58	78	PPV	100
Grantz et al†	**SGA**	349	43	0	K	0.67 (0.64-0.69)					
	AGA	123	3661	115	Sens	65.7					
	LGA	0	225	277	PPV	84.2					
Shivkumar et al§	SGA	406	253	0	K	0.59 (0.56-0.62)	100	24	0	K	0.77 (0.72-0.82)
	AGA	8	2946	254	Sens	98.1	8	843	37	Sens	92.6
	LGA	0	0	72	PPV	61.6	0	11	63	PPV	80.6
Stirrup et al†	**SGA**	296	27	0	K	0.69 (0.66-0.72)	170	0	0	K	0.54 (0.51-0.57)
	AGA	108	3825	215	Sens	73.3	234	3713	182	Sens	42.1
	LGA	0	68	254	PPV	91.6	0	207	287	PPV	100

		Dichorionic					Monochorionic-Diamniotic				
		Our cohort					Our cohort				
Singleton Standards		**SGA**	**AGA**	**LGA**		**Correlation**	**SGA**	**AGA**	**LGA**		**Correlation**
INTERGROWTH-21 st †	**SGA**	308	112	0	K	0.59 (0.56-0.62)	86	56	0	K	0.56 (0.48-0.62)
	AGA	106	2820	86	Sens	74.4	22	751	31	Sens	79.6
	LGA	0	267	240	PPV	73.3	0	71	69	PPV	60.6
Savirón-Cornudella et al †	**SGA**	465	576	0	K	0.56 (0.54-0.59)	140	232	0	K	0.49 (0.44-0.54)
	AGA	7	3271	144	Sens	98.5	2	849	42	Sens	100
	LGA	0	82	248	PPV	44.7	0	31	74	PPV	37.6
WHO†	**SGA**	424	396	0	K	0.59 (0.56-0.61)	118	145	0	K	0.50 (0.45-0.55)
	AGA	48	3384	135	Sens	89.8	22	905	40	Sens	84.3
	LGA	0	149	257	PPV	51.7	0	62	76	PPV	44.9

* From 23th to 38th week of gestational age.
† From 17th to 38th week of gestational age.
‡ From 24th to 38th week of gestational age.
§ From 22nd week to 38th week of Gestational age. K: Cohen's kappa coefficient. PPV: positive predictive value of each standard to predict small for gestational age cases provided by our standard. Sens: sensitivity of each standard to predict small for gestational age cases provided by our standard.

kept in mind for the clinical use of the weight percentiles estimated by ultrasound.

Future research should consider the assessment of maternal nutrition and body weight gain during twin pregnancies since it may influence fetal growth. Women with twins have increased metabolic demands, different degree of physical activity and APOs. For instance, in singleton pregnancies increased BMI is associated with preterm birth while in healthy twin gestations preterm birth is associated with maternal undernutrition [32]. In addition, BMI is not an appropriate indicator of weight and of feto-maternal nutrition during pregnancy. On the other hand, in twin pregnancies, excessive gestational weight gain is associated with higher rates of preeclampsia and gestational diabetes and associated complications [33]. Despite this, even the strict application of the recommendations of the Institute of Medicine does no avoid possible pregnancy complications [34]. In our cohort, the BMIs were similar in the two subgroups of pregnant women. Future work should identify the appropriate weight gain that may influence fetal growth in twin pregnancies.

As a limitation, the MCDA study was based on 83 mothers, although the 1455 weights used in the building model were large enough, our study has a smaller sample size than previous studies.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or non-for-profit sectors.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgements

We gratefully thank all the staff of the Ultrasound and Prenatal Diagnosis Unit of the Miguel Servet University Hospital (Zaragoza, Spain) for their help in the acquisition and management of data.

References

[1] Peristat Euro, Macfarlane AJ. Euro-Peristat Project. European Perinatal Health Report. Core indicators of the health and care of pregnant women and babies in Europe in 2015. Euro-Peristat; 2016. . (1 November 2018) http://www.europeristat.com.
[2] Laine K, Murzakanova G, Sole KB, Pay AD, Heradstveit S, Räisänen S. Prevalence and risk of pre-eclampsia and gestational hypertension in twin pregnancies: a population-based register study. BMJ Open 2019;9(7)e029908.
[3] Hadlock FP, Harrist RB, Sharman RS, Deter RL, Park SK. Estimation of fetal weight with the use of head, body, and femur measurements: a prospective study. Am J Obstet Gynecol 1985;151(3):333–7.
[4] Ananth CV, Vintzileos AM, Shen-Schwarz S, Smulian JC, Lai YL. Standards of birth weight in twin gestations stratified by placental chorionicity. Obstet Gynecol 1998;91(6):917–24.
[5] Liao AW, Brizot Mde L, Kang HJ, Assunção RA, Zugaib M. Longitudinal reference ranges for fetal ultrasound biometry in twin pregnancies. Clinics (Sao Paulo) 2012;67(5):451–5.
[6] Araujo Júnior E, Ruano R, Javadian P, Martins WP, Elito Jr J, Pires CR, et al. Reference charts for fetal biometric parameters in twin pregnancies according to chorionicity. Prenat Diagn 2014;34(4):382–8.
[7] Shivkumar S, Himes KP, Hutcheon JA, Platt RW. An ultrasound-based fetal weight reference for twins. Am J Obstet Gynecol 2015;213(2)224 e1-9.
[8] Grantz KL, Grewal J, Albert PS, Wapner R, D'Alton ME, Sciscione A, et al. Dichorionic twin trajectories: the NICHD Fetal Growth Studies. Am J Obstet Gynecol 2016;215(2)221 e1-221.e16.
[9] Gabbay-Benziv R, Crimmins S, Contag SA. Reference Values for Sonographically Estimated Fetal Weight in Twin Gestations Stratified by Chorionicity: A Single Center Study. J Ultrasound Med 2017;36(4):793–8.
[10] Stirrup OT, Khalil A, D'Antonio F, Thilaganathan B, Southwest Thames Obstetric Research Collaborative (STORK). Fetal growth reference ranges in twin pregnancy: analysis of the Southwest Thames Obstetric Research Collaborative (STORK) multiple pregnancy cohort. Ultrasound in Obstetrics & Gynecology 2015;45(3):301–7.
[11] Odibo AO, Cahill AG, Goetzinger KR, Harper LM, Tuuli MG, Macones GA. Customized growth charts for twin gestations to optimize identification of small-for-gestational age fetuses at risk of intrauterine fetal death. Ultrasound Obstet Gynecol 2013;41(6):637–42.
[12] Kalafat E, Sebghati M, Thilaganathan B, Khalil A, Southwest Thames Obstetric Research Collaborative (STORK). Predictive accuracy of Southwest Thames Obstetric Research Collaborative (STORK) chorionicity-specific twin growth charts for stillbirth: a validation study. Ultrasound in Obstetrics & Gynecology 2019;53(2):193–9.
[13] Joseph KS, Fahey J, Platt RW, Liston RM, Lee SK, Sauve R, et al. An outcome-based approach for the creation of fetal growth standards: do singletons and twins need separate standards? Am J Epidemiol 2009;169(5):616–24.
[14] Papageorghiou AT, Bakoulas V, Sebire NJ, Nicolaides KH. Intrauterine growth in multiple pregnancies in relation to fetal number, chorionicity and gestational age. Ultrasound Obstet Gynecol 2008;32(7):890–3.
[15] D'Antonio F, Odibo A, Berghella V, Khalil A, Hack K, Saccone G, et al. Perinatal mortality, timing of delivery and prenatal management of monoamniotic twin pregnancy: systematic review and meta-analysis. Ultrasound in Obstetrics & Gynecology 2019;53(2):166–74.
[16] Committee Opinion No 700. Methods for Estimating the Due Date. Obstet Gynecol 2017;129(5):e150–4.
[17] Stirnemann J, Villar J, Salomon LJ, Ohuma E, Ruyan P, Altman DG, et al. International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st). International estimated fetal weight standards of the INTERGROWTH-21st project. Ultrasound Obstet Gynecol 2017;49(4):478–86.
[18] Laird N, Ware J. Random-Effects Models for Longitudinal Data. Biometrics 1982;38(4):963–74.
[19] Ohuma EO, Altman DG. International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st Project). Statistical methodology for constructing gestational age-related charts using cross-sectional and longitudinal data: The INTERGROWTH-21st project as a case study. Statistics in medicine 2019;38(19):3507–26.
[20] Hadlock FP, Harrist RB, Martinez-Poyer J. In utero analysis of fetal growth: a sonographic weight standard. Radiology 1991;181(1):129–33.
[21] Harrell FE. Regression Modeling Strategies: with applications to linear models, logistic regression, and survival analysis. New York, USA: Springer-Verlag New York, Inc.; 2010.
[22] Nakagawa S, Schielzeth H. A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods in ecology and evolution 2013;4(2):133–42.
[23] Saviron-Cornudella R, Esteban LM, Lerma D, Cotaina L, Borque A, Sanz G, et al. Comparison of fetal weight distribution improved by paternal height by Spanish standard versus INTERGROWTH 21st standard. J Perinat Med 2018;46(7):750–9.
[24] Kiserud T, Piaggio G, Carroli G, Widmer M, Carvalho J, Neerup Jensen L, et al. The World Health Organization Fetal Growth Charts: A Multinational Longitudinal Study of Ultrasound Biometric Measurements and Estimated Fetal Weight. PLoS Medicine 2017;14(1):e1002220.
[25] R Core Team. R: A languaje and environment for statistical computing. Vienna, Austria: R Foundation [56_TD$DIFF]for Statistical Computing; 2018. . (1 January 2019) http://www.R-project.org.
[26] Torres X, Bennasar M, Eixarch E, Rueda C, Goncé A, Muñoz M, et al. Gender-Specific Antenatal Growth Reference Charts in Monochorionic Twins. Fetal Diagn Ther 2018;44(3):202–9.
[27] Ghi T, Prefumo F, Fichera A, Lanna M, Periti E, Persico N, Am J Obstet Gynecol 2017;216(5)514 e1-514.e17.
[28] Hutcheon JA, Zhang X, Platt RW, Cnattingius S, Kramer MS. The case against customised birthweight standards. Paediatr Perinat Epidemiol 2011;25(1):11–6.
[29] Carberry AE, Gordon A, Bond DM, Hyett J, Raynes-Greenow CH, Jeffery HE. Customised versus population-based growth charts as a screening tool for detecting small for gestational age infants in low-risk pregnant women. Cochrane Database Syst Rev 2014;16(5):CD008549.
[30] Milner J, Arezina J. The accuracy of ultrasound estimation of fetal weight in comparison to birth weight: A systematic review. Ultrasound 2018;26(1):32–41.
[31] Khalil A, D'Antonio F, Dias T, Cooper D, Thilaganathan B, Southwest Thames Obstetric Research Collaborative (STORK). Ultrasound estimation of birth weight in twin pregnancy: comparison of biometry algorithms in the STORK multiple pregnancy cohort. Ultrasound Obstet Gynecol 2014;44:210–20.
[32] Ram M, Berger H, Lipworth H, Geary M, McDonald SD, Murray-Davis B, et al. DOH-Net (Diabetes, Obesity and Hypertension in Pregnancy Research Network) and SOON (Southern Ontario Obstetrical Network) Investigators. The relationship between maternal body mass index and pregnancy outcomes in twin compared with singleton pregnancies. Int J Obes 2020;44(1):33–44, doi:http://dx.doi.org/10.1038/s41366-019-0362-8.
[33] Pécheux O, Garabedian C, Drumez E, Mizrahi S, Cordiez S, Deltombe S, et al. Maternal and neonatal outcomes according to gestational weight gain in twin pregnancies: Are the Institute of Medicine guidelines associated with better outcomes? Eur J Obstet Gynecol Reprod Biol 2019;234:190–4.
[34] Wang L, Wen L, Zheng Y, Zhou W, Mei L, Li H, et al. Association Between Gestational Weight Gain and Pregnancy Complications or Adverse Delivery Outcomes in Chinese Han Dichorionic Twin Pregnancies: Validation of the Institute of Medicine (IOM) 2009 Guidelines. Med Sci Monit 2018;(24):8342–7.

7.3. Valor predictivo del score de riesgo genético en el desarrollo de adenomas colorrectales

El estudio [34] que se presenta en esta sección se centra en la evaluación de factores de riesgo en el desarrollo de adenomas colorrectales. En concreto, la investigación se enfoca en evaluar la capacidad predictiva del score de riesgo genético, un aspecto que no ha sido suficientemente explorado en este contexto.

Diversas variantes genéticas, como los polimorfismos de un solo nucleótido (SNP, siglas en inglés), han sido identificadas en numerosos estudios con el riesgo de cáncer colorrectal. Sin embargo, por sí solas, no ofrecen un riesgo clínicamente relevante. La combinación de estos marcadores genéticos, conocido como score de riesgo genético (GRS, siglas en inglés), es una práctica común que puede proporcionar una capacidad predictiva superior [20]. Aunque existen diversos modelos genéticos, el modelo aditivo, que suma los alelos de riesgo de cada SNP, es el más común en la construcción del GRS.

En nuestro estudio consideramos tanto enfoques ponderados (weighted GRS) como no ponderados (unweighted GRS). El GRS no ponderado asume que todos los SNPs tienen el mismo efecto y se obtiene como la suma de todos los alelos de riesgo de los SNPs: unweighted GRS $= \sum_{i=1}^{n} r_i$, donde r_i representa el número de alelos de riesgo que tiene el paciente en el i-ésimo SNP. Por otro lado, el GRS ponderado asume un efecto diferente para cada SNP y se obtiene como la suma de todos los alelos de riesgo multiplicados por un peso w_i: weighted GRS $= \sum_{i=1}^{n} w_i r_i$. Para la contrucción de estos scores genéticos, se utilizó la regresión logística considerando la información genética del paciente. El coeficiente estimado para cada SNP es el que se consideró como peso w_i.

Se aplicaron test de comparación de curvas ROC para evaluar la diferencia discriminatoria entre los modelos. Nuestros hallazgos mostraron que, además del sexo y la edad, el GRS es un factor de riesgo importante para el desarrollo de adenomas colorrectales. En resumen, el uso de modelos lineales permitió evaluar el factor de riesgo genético en el desarrollo de adenomas colorrectales, proporcionando un puntaje (GRS) que es facilmente interpretable por el clínico. Estos resultados pueden influir en la práctica clínica, donde los GRS pueden ser utilizados para estratificar el riesgo de cáncer colorrectal o en programas de detección, mejorando su precisión.

A continuación se presenta el artículo, donde se proporciona información más detallada de este estudio.

Digestive Diseases and Sciences
https://doi.org/10.1007/s10620-021-07218-5

ORIGINAL ARTICLE

Predictive Value of Genetic Risk Scores in the Development of Colorectal Adenomas

Carla J. Gargallo-Puyuelo[1,2] · Rocío Aznar-Gimeno[3] · Patricia Carrera-Lasfuentes[2,4] · Ángel Lanas[1,2,4,5] · Ángel Ferrández[1,2] · Enrique Quintero[6,7] · Marta Carrillo[6] · Inmaculada Alonso-Abreu[6] · Luis M. Esteban[8] · María de la Vega Rodrigálvarez-Chamarro[3] · Rafael del Hoyo-Alonso[3] · María Asunción García-González[2,4,9]

Received: 11 June 2021 / Accepted: 2 August 2021

Abstract

Introduction Unlike colorectal cancer (CRC), few studies have explored the predictive value of genetic risk scores (GRS) in the development of colorectal adenomas (CRA), either alone or in combination with other demographic and clinical factors.

Methods In this study, genomic DNA from 613 Spanish Caucasian patients with CRA and 829 polyp-free individuals was genotyped for 88 single-nucleotide polymorphisms (SNPs) associated with CRC risk using the MassArray™ (Sequenom) platform. After applying a multivariate logistic regression model, five SNPs were selected to calculate the GRS. Regression models adjusted by sex, age, family history of CRC, chronic use of NSAIDs, low-dose ASA, and consumption of tobacco were built in order to study the association between GRS and CRA risk. We evaluated the discriminatory capacity using the area under the receiver operating characteristic curve (AUC). The interactions between demographic information and GRS were also analyzed.

Results Significant associations between high GRS values and risk of CRA for analyzed models were observed. In particular, patients with higher GRS values had 2.3–2.6-fold increase in risk of CRA compared to patients with middle values. Combining sex and age with the GRS significantly increased the discriminatory accuracy of the univariate model with GRS alone. The best model achieved an AUC value of 0.665 (95% CI: 0.63–0.69). The GRS showed a different behavior depending on sex and age.

Conclusion Our findings showed that, besides sex and age, GRS is an important risk factor for development of CRA and may be useful for CRC risk stratification and adaptation of screening programs.

Keywords Genetic risk scores · Colorectal adenomas · Colorectal screening · Predictive value · Single-nucleotide polymorphism

Carla J. Gargallo-Puyuelo y Rocío Aznar-Gimeno is dual first authorship due to equal contribution.

✉ Carla J. Gargallo-Puyuelo
carlajerusalen@hotmail.com

1 Department of Gastroenterology, Hospital Clínico Universitario Lozano Blesa, Av: San Juan Bosco, no 15. PC, 50009 Zaragoza, Spain

2 Instituto de Investigación Sanitaria Aragón (IIS Aragón), 50009 Zaragoza, Spain

3 Instituto Tecnológico de Aragón (ITAINNOVA), 50018 Zaragoza, Spain

4 CIBERehd, Zaragoza, Spain

5 School of Medicine, University of Zaragoza, 50009 Zaragoza, Spain

6 Hospital Universitario de Canarias, Santa Cruz de Tenerife, Spain

7 School of Medicine, University of La Laguna, Canary Islands, Santa Cruz de Tenerife, Spain

8 Department of Applied Mathematics, Escuela Universitaria Politécnica de La Almunia, Universidad de Zaragoza, 50100 Zaragoza, Spain

9 Instituto Aragonés de Ciencias de La Salud (IACS), 50009 Zaragoza, Spain

Published online: 13 August 2021

Introduction

Colorectal cancer (CRC) is the third most frequent cancer worldwide with 1,849,518 (10.2%) new cases diagnosed in 2018. Moreover, CRC represents the second most common cause of cancer death with 880,792 deaths occurred in that year [1]. Most CRCs develop from premalignant lesions (mainly adenomas) that take years to transform into a cancer. This lapse of time can allow for early detection of lesions through CRC screening programs developed to reduce both, incidence and mortality rates.

It is now well accepted that CRC results from complex interactions between environmental and genetic factors. In the last two decades, numerous genome-wide association studies (GWAS) on CRC risk have identified more than 80 common single-nucleotide polymorphisms (SNPs) in low penetrance genes [2–4]. These genetic variants do not provide clinically relevant risk by themselves, but combination of risk alleles in a polygenic model has been reported to increase the risk of CRC in an additive or exponential way. In fact, the elaboration of genetic risk scores (GRS), combining multiple SNPs associated with CRC, is becoming a very attractive issue showing very promising results for a more accurate CRC risk stratification. Recent studies have suggested that adding genetic and environmental information into a CRC risk prediction model may significantly increase the discriminatory accuracy over current screening models based mainly on age and family history of CRC [4–9].

Unlike CRC, very few studies [6] have explored the predictive value of GRS in the development of premalignant colorectal lesions, namely colorectal adenomas. Indeed, performance of risk prediction models for early events of the adenoma-carcinoma sequence may offer the greatest potential benefit for CRC prevention. Trying to address this specific issue, the aim of our study was to evaluate the predictive capacity of GRS on the risk of colorectal adenomas, either alone or in combination with other clinical and demographic characteristics.

Material and Methods

Study Population

This study included 1,500 patients who were scheduled for colonoscopy, either by symptoms or through the CRC screening programs (in the average-risk population and in first-degree relatives of patients with nonsyndromic CRC), at two general Spanish hospitals from May 2010 to May 2014. The study design has been previously described in detail [10]. Criteria for exclusion included hereditary CRC syndromes, a personal history of CRC or inflammatory bowel disease, previous polypectomy, age < 18 years old and ethnicity other than Caucasian.

All subjects underwent at least one colonoscopy. The quality of the bowel preparation for colonoscopy, evaluated by the Boston Bowel Preparation Scale, was good or very good in most cases (87.3%). Similarly, the rate of complete colonoscopies was high (97.7%). Information concerning demographic and clinical characteristics such as family history of CRC (any reported CRC in FDR or two or more CRC cases in second-degree relatives), smoking habit, and chronic use of nonsteroidal anti-inflammatory drugs (NSAIDs) or acetylsalicylic acid (ASA) were considered as previously described [10]. Because all clinical variables showed very low missing rates (< 7%), the missing values were replaced by the most frequent value observed within each one.

After completion of the interview, 10 ml of peripheral blood was obtained from each individual and collected into EDTA (ethylenediaminetetraacetic acid) tubes for subsequent DNA extraction and genotyping. Once processed, whole blood samples were aliquoted and stored at − 80 °C until analyzed.

All subjects gave written informed consent to the study which was approved by the institutional ethic committee of each participating hospital (University Hospital Lozano Blesa of Zaragoza, and University Hospital of the Canary Islands in Tenerife). The study has been performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments.

Genotyping

Genomic DNA was extracted from EDTA-preserved whole blood using the automated DNA isolation system AutoGen-Flex 3000. As formerly described in our study by Gargallo et al. [10], a total of 1,500 patients were genotyped for a panel of 88 SNPs previously known to confer genetic susceptibility to CRC. Genotyping was performed at the Spanish National Genotyping Centre (CEGEN-Santiago de Compostela) by using the Sequenom MassARRAY iPLEX platform. Eleven out of the 88 SNPs analyzed in the study showed statistically significant associations with the risk of colorectal adenomas (False Discovery Rate < 0.05). Therefore, we selected these 11 SNPs (rs10505477, rs10795668, rs11255841, rs13181, rs16260, rs1728785, rs4779584, rs647161, rs6983267, rs8180040, and rs9929218) for subsequent analyses to develop a risk prediction model of colorectal adenomas. The general characteristics of these SNPs are shown in Table 1.

Table 1 General characteristics of the SNPs analyzed in the study

Db SNP ID[a]	Mapped Gen	Chr	Position[b]	SNP type[c]	A/a[d]	MAF[e]
rs8180040	*ND*	3	47,347,457	ND	T/A	0.25
rs647161	*C5orf66*	5	135,163,402	Intronic	A/C	0.46
rs10505477	*CASC8*	8	127,395,198	Intronic	A/G	0.42
rs6983267	*CASC8*	8	127,401,060	Intronic	G/T	0.39
rs10795668	*LINC0079*	10	8,659,256	Intronic	G/A	0.23
rs11255841	*LINC0079*	10	8,697,617	Intronic	T/A	0.25
rs4779584	*GREM1-SCG5*	15	32,702,555	Intergenic	C/T	0.49
rs1728785	*ZFP90*	16	68,557,327	Intronic	C/A	0.18
rs16260	*CDH1*	16	68,737,131	upstream variant	C/A	0.28
rs9929218	*CDH1*	16	68,787,043	Intronic	G/A	0.25
rs13181	*ERCC2*	19	45,351,661	K751Q	T/G	0.24

Chr, chromosome number. ND, not described

[a]SNP identification according to the NCBI data base.

[b]Chromosome position according to the Genome Reference Consortium Human Build 38.p2 (GRCh38.p2).

[c]Location in the gen.

[d]Major/Minor alleles.

[e]MAF: Minor allele frequency for European population by free access database [dbSNP (National Center for Biotechnology Information, NCBI; http://www.ncbi.nlm.nih.gov/snp/)]

From the initial 1,500 recruited subjects, 1,442 patients (613 patients with colorectal adenomas and 829 polyp-free individuals) were successfully genotyped for the 11 mentioned SNPs and included in the study for final analysis.

Statistical Analyses

An initial descriptive exploratory analysis of all clinical variables was carried out. Continuous variables were expressed as mean with standard deviation (SD) and as median with interquartile range (IQR), whereas qualitative variables were expressed as frequencies and percentages. Differences between populations with and without colorectal adenomas were evaluated with Chi-square ($\chi 2$) test for qualitative variables and with Mann–Whitney test for continuous variables. Normality was tested using the Shapiro–Wilk test. A step-by-step multivariate logistic regression model was implemented among the 11 SNPs previously mentioned. Finally, five SNPs (rs10505477, rs11255841, rs13181, rs4779584, and rs8180040) were selected in order to calculate the genetic risk score (GRS). The number of risk alleles was coded as 0, 1 or 2 for each SNP assuming a log-additive genetic effect. The methods to compute GRS were based in both, weighted and unweighted ways. The risk alleles were determined by the estimated coefficients of the logistic regression. The unweighted GRS assumes that all SNPs have the same effect and was obtained as the sum of all risk alleles of the five selected SNPs. However, the weighted GRS assumes a different effect for each SNP and it was obtained as the sum of all risk alleles multiplied by their respective coefficients of the logistic regression models. As weighted and unweighted GRS yielded very similar results, only results for the unweighted GRS are presented in the main text. (Analyses for weighted GRS are summarized in the Supplementary Material.)

The distribution of the GRS in patients with and without adenomas was also analyzed. In order to study the association between the GRS and risk of adenomas, logistic regression models were constructed and odds ratios (ORs) and 95% confidence intervals (CIs) were calculated. Multivariate models were adjusted by all variables retrieved in the study (sex, age, family history, chronic use of NSAIDs or low-dose ASA, and consumption of tobacco) and for only those ones statistically significant. The interactions between covariates and GRS were also analyzed.

The discriminatory accuracy of models was evaluated using the area under the receiver operating characteristics (ROC) curve (AUC). To avoid an overestimation of the value, the reported AUC value was internally calculated using a fivefold cross-validation and the 95% CI were obtained using 2,000 stratified bootstrap replicates. The test of comparison of ROC curves proposed by Pepe et al. [11] was applied to assess the discriminatory difference between models.

The level of significance in the study was established at 0.05. Analyses were performed using the R v.3.5.3 programming language (The R Foundation for statistical computing, Vienna, Austria) [12]. In particular, the Predict ABEL R package was used to compute the GRS [13].

Table 2 Clinical and demographic characteristics of the study population

Characteristics		Total $N=1442$	Without adenomas $N=829$	With adenomas $N=613$	p value
Age (years)	Mean (SD)	54.55 (9.46)	53.52 (10.09)	55.94 (8.34)	**< 0.001**
	Median (IQR)	55 (49–61)	54 (48–61)	57 (50–62)	
Sex	Female	749 (51.94%)	498 (60.07%)	251 (40.95%)	**< 0.001**
Smoking	Yes	373 (25.87%)	207 (24.97%)	166 (27.08%)	0.061
	No	795 (55.13%)	478 (57.66%)	317 (51.71%)	
	Former smoker	274 (19.00%)	144 (17.37%)	130 (21.20%)	
Chronic NSAIDs	Yes	93 (6.45%)	61 (7.36%)	32 (5.22%)	0.127
Chronic low-dose ASA	Yes	80 (5.55%)	43 (5.29%)	37 (6.06%)	0.562
Family history	Yes	721 (50.00%)	414 (49.94%)	307 (50.08%)	1.000

Bold indicates p value < 0.05

N number of individuals. *SD* standard deviation. *IQR* interquartile range. *NSAIDs* nonsteroidal anti-inflammatory drugs. *ASA* acetylsalicylic acid

Results

Clinical and demographic characteristics of the study population are shown in Table 2. Some statistically significant differences between patients with and without adenomas were observed. In this context, patients with adenomas were significantly older (mean 55.94 vs. mean 53.5) and showed predominance for male gender (59.05% vs. 39.93%) than patients without adenomas. However, no significant differences were observed regarding consumption of tobacco and chronic use of NSAIDs or low-dose ASA. The proportion 50–50 of family history was remained in both groups, in accordance with the study design described in Gargallo et al. [10].

The degree of association between each of the five SNPs (rs10505477, rs11255841, rs13181, rs4779584, and rs8180040) selected to build the GRS and risk of colorectal adenoma is represented by means of ORs and their 95% CIs in Fig. 1. The observed associations for all SNPs in our study were in the same direction as reported previously in their respective discovery studies. With the exception of the rs4779584 SNP, major alleles (most frequent alleles) behaved as risk alleles for colorectal adenomas.

The mean number of risk alleles was significantly different between subjects with (5.659 ± 1.494) and without (5.292 ± 1.473) adenomas (difference: 0.367 alleles, 95% CI 0.212–0.522, $p < 0.001$). This trend is shown through the distribution of the GRS values for patients with and without adenomas in Fig. 2. Both populations follow a normal distribution with a maximum value of five risk alleles for the population without adenomas and six in the population with them, value from which a change occurs toward a higher increasing percentage of high GRS values in the population with adenomas with respect to the population without them. The GRS was categorized into six levels of genetic risk: ≤ 3, 4, 5, 6, 7, ≥ 8 risk alleles, due to the low representativeness in the extremes values. Five risk alleles were considered as reference in the analysis since it was the median value observed. A similar trend was observed when weighted GRS considering quartiles was performed (Supplementary Fig. 1).

The association between unweighted GRS and risk of colorectal adenomas is shown in Table 3. Four different models were constructed: (1) univariate GRS model; (2) multivariate model adjusted by the statistically significant variables sex and age; (3) multivariate model adjusted by the pairwise interaction between sex and age; and (4) multivariate model adjusted by all variables retrieved in the study (sex, age, consumption of tobacco and chronic use of NSAIDs and low-dose ASA). Significant associations between high GRS values and risk of colorectal adenomas were found for all analyzed models. In particular, individuals with ≥ 8 risk alleles showed a significant increase in adenoma risk compared with subjects with GRS of 5. Values ranged from 2.3-folds (95% CI 1.63–3.32) in the univariate model to nearly 2.6-folds (95% CI 1.8–3.75) in the multivariate model adjusted by the interaction between sex and age.

The analysis using weighted GRS reported similar associations (Supplementary Table 1). The ORs per quartiles of the weighted GRS are presented in the table in which significant associations between high GRS values and risk of colorectal adenomas were also found.

The ROC curves of the models analyzed are displayed in Fig. 3. We found that combining sex and age to the univariate GRS prediction model significantly increased discriminatory accuracy from 0.571 (95% CI: 0.54–0.60) to 0.655 (95% CI: 0.62–0.68). In fact, the highest prediction values were observed in the combined GRS-interaction sex and age model (AUC: 0.665; 95% CI: 0.63–0.69). However, no significant improvement was observed when considering in the model the other variables retrieved in the study (AUC: 0.649; 95% CI: 0.62–0.68). The test also showed a

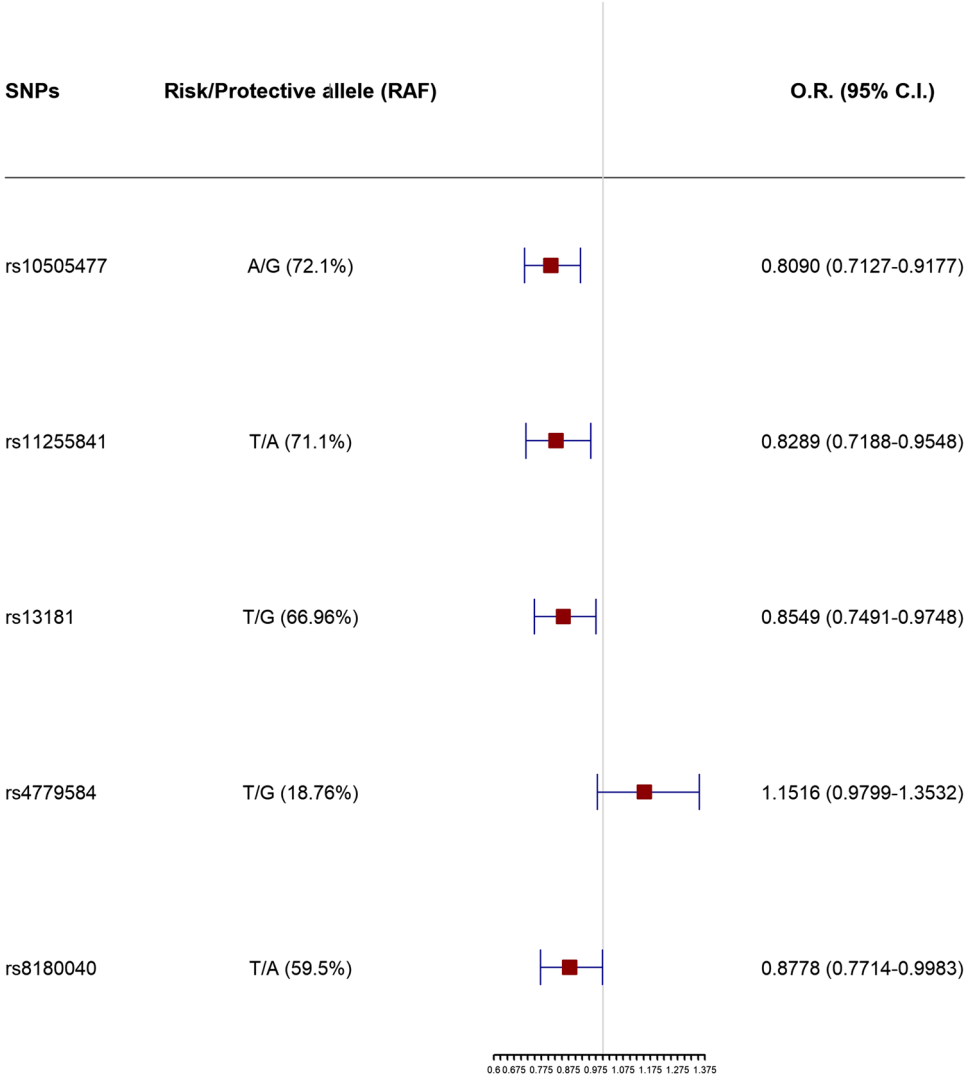


Fig. 1 Association between the five selected SNPs and colorectal adenoma risk. RAF: risk allele frequency. The coding of each SNP of the multivariate model represented was as follows: the value 1 was assigned to the most frequent allele and value 0 to the less frequent allele

significant predictive improvement by including the GRS in the model adjusted for sex and age.

Interactions between the GRS and the variables sex and age are shown in Fig. 4. Differences in the effect of the GRS according to sex and age are shown, pointing out to an association between higher GRS values in men and individuals ≥ 50 years. Association between adenomas risk and combination of GRS and age and GRS and sex is shown in Table 4. Due to the results obtained and being the cutoff point established in the general population for CRC screening, 50 years was the cutoff point chosen for the age categorization. The patients with the lower predisposition of adenomas (women with the lowest GRS and people < 51 years with the lowest GRS) were taken as references. The highest ORs were reached in men (OR: 4.58; 95% CI: 2.43–8.8; $p < 0.001$) and individuals older than 50 years (OR: 6.74; 95% CI: 3.25–14.72; $p < 0.001$) with GRS values ≥ 8.

Discussion

CRC is a very common cancer worldwide but it is also one of the most preventable and curable cancer if early detected. Current clinical guidelines recommend CRC screening based mainly on age and family history of CRC [14, 15]. Risk prediction models are very useful tools to more accurately define low- and high-risk individuals, which is essential in precision medicine.

Most of CRCs develop from adenomas through the adenoma–carcinoma sequence. Previous evidence indicates that many of the SNPs reported in the literature to

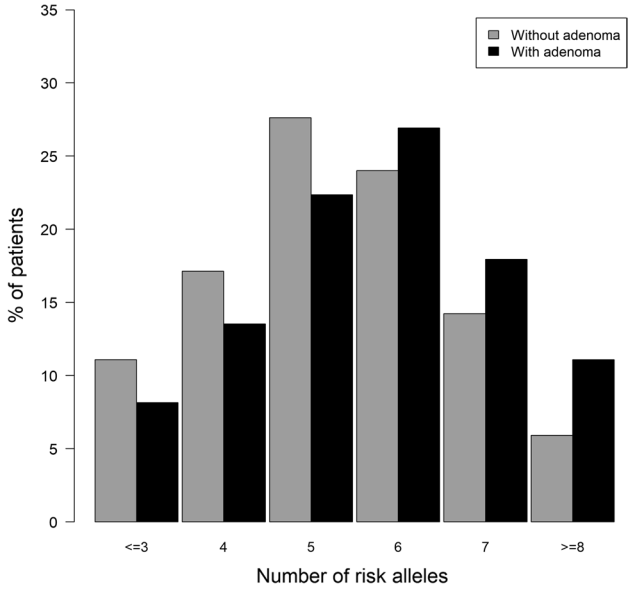


Fig. 2 Distribution of unweighted GRS values in patients with and without adenomas

be associated with CRC are also related to higher risk of adenomas due to their involvement in early carcinogenesis [10, 16]. The main objective of this work was to explore the predictive value of GRS in the development of colorectal adenomas, either alone or in combination with other clinical and demographic characteristics. After evaluating 88 SNPs previously associated with CRC or colorectal adenomas [10], we included in our study a GRS based on five SNPs. Selected SNPs were the following: rs10505477, located in the long non coding *CASC8* gene (cancer susceptibility candidate 8), rs11255841 in LINC00709 (long intergenic non protein coding RNA 709), rs13181in the *ERCC2* (excision repair 2) gene, rs4779584 located between the *GREM1* (gremlin 1) and *SCG5* (secreto granin V) genes and rs8180040 in the *PTPN23* (protein tyrosine phosphatase non-receptor type 23) gene. In agreement with their respective discovery studies, major alleles (most frequent alleles) behaved as risk alleles for colorectal adenomas in our study with the exception of the rs4779584 SNP.

In addition to genetic variants, most published risk prediction models of CRC include environmental/lifestyle factors and/or family history of CRC. The number of SNPs used to construct the GRS in the different studies ranges from 10 to 90, reporting an increased number of SNPs in the most recent studies. Dunlop et al. [17] designed a CRC predictive model based on 10 SNPs, age, gender, and family history of CRC, obtaining an AUC value of 0.59. Yarnall et al. [18] reported a slightly better AUC value (0.61) combining a GRS based on 14 SNPs and alcohol intake. Ibáñez-Sanz et al. [9] reported a discriminatory accuracy value of 0.63 for their combining modifiable risk factors, family history, and a GRS (21 SNPs). More recently, Weigl et al. [6] built a GRS based on 48 SNPs. They observed that participants in the upper tertile of the GRS almost tripled the risk of advanced neoplasms compared to patients in the

Table 3 Association between unweighted GRS and risk of adenomas

GRS values	Model 1[a] OR (95% CI) *p* value	Model 2[b] OR (95% CI) *p* value	Model 3[c] OR (95% CI) *p* value	Model 4[d] OR (95% CI) *p* value
≤3	0.908 (0.645–1.273) 0.642	0.854 (0.601–1.209) 0.459	0.861 (0.603–1.222) 0.484	0.876 (0.615–1.241) 0.534
4	0.977 (0.731–1.303) 0.895	0.968 (0.718–1.303) 0.858	0.975 (0.722–1.315) 0.889	0.976 (0.723–1.314) 0.891
5	**Ref**	**Ref**	**Ref**	**Ref**
6	1.386 (1.082–1.777) **0.031**	1.431 (1.108–1.849) **0.021**	1.460 (1.129–1.890) **0.016**	1.443 (1.117–1.866) **0.019**
7	1.558 (1.177–2.065) **0.010**	1.622 (1.213–2.169) **0.006**	1.650 (1.233–2.210) **0.005**	1.642 (1.228–2.199) **0.019**
> =8	2.320 (1.628–3.320) **< 0.001**	2.500 (1.737–3.617) **< 0.001**	2.592 (1.799–3.752) **< 0.001**	2.521 (1.749–3.650) **< 0.001**
Continuous Value	1.201 (1.127–1.281) **< 0.001**	1.230 (1.151–1.314) **< 0.001**	1.237 (1.158–1.323) **< 0.001**	1.229 (1.150–1.313) **< 0.001**

Bold indicates *p* value < 0.05

[a]Univariate model.

[b]Multivariate model adjusted by sex and age.

[c]Multivariate model adjusted by the interaction between sex and age.

[d]Multivariate model adjusted by sex, age, family history of CRC, NSAIDs chronic consumption, aspirin chronic consumption and tobacco use.

CI confidence interval; *GRS* genetic risk score; *OR* odds ratio; *Ref* reference

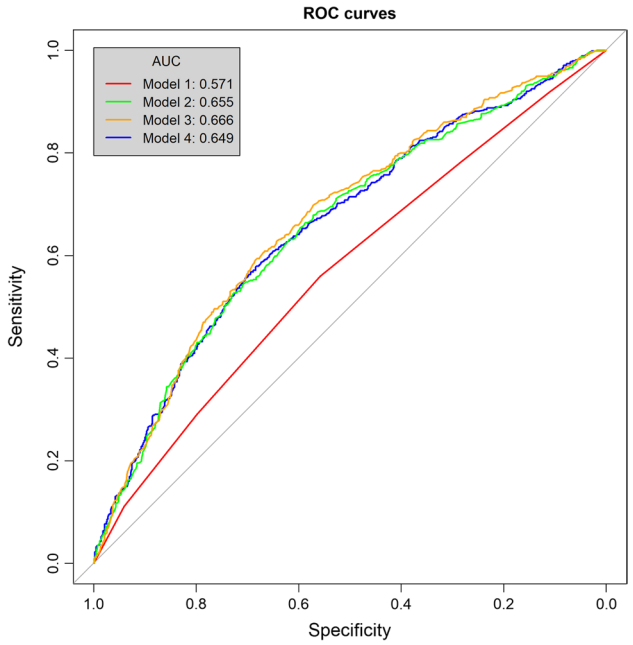


Fig. 3 ROC curves of the models analyzed. Model 1: Univariate model (unweighted GRS). Model 2: Multivariate model adjusted by sex and age. Model 3: Multivariate model adjusted by the interaction between sex and age. Model 4: Multivariate model adjusted by sex, age, family history of CRC, NSAIDs chronic consumption, aspirin chronic consumption, and tobacco use

lower tertile. An increasing number of SNPs (63) were evaluated by Jeon et al. [5] along with family history data and 19 lifestyle and environmental factors. Their model combining all risk factors estimated CRC risk with a discriminatory accuracy value of 0.63 for men and 0.62 for women. Similar prediction values for advanced colorectal neoplasm were obtained by Balavarca et al. [19] with combined environmental-genetic score model (AUC = 0.63).

Unlike CRC, very few studies have explored the predictive value of GRS in the development of premalignant colorectal lesions (namely, colorectal adenomas) [6, 16, 20]. In our study, the best predictive accuracy for the four models evaluated has an AUC value of 0.66 and belongs to the model that considers the GRS along with the variables sex and age. Current evidence shows that age and sex are the most important risk factors for developing CRC and/or adenomas along with family history. In fact, some studies suggest that CRC screening programs should be started at an earlier age in men than in women [21–23]. In accordance with these recommendations, our analysis shows that high values in GRS carry more risk of adenomas in men and individuals older than 50 years. This is in line with the results reported by Weigl et al. [6] showing a joint relationship between age and GRS. According to the authors, the risk of the 60-year-old medium GRS group was reached in the low GRS group at age of 73 years, and at age of 56 in the high GRS group. In our study, being a man also increases considerably the adenoma risk carry by high values of GRS. Only women with ≥ 8 risk alleles have higher risk of adenomas than men with ≤ 4 risk alleles. Currently, most countries recommend starting CRC screening programs at age of 50 in average-risk population [14, 15]. However, our findings support the hypothesis that CRC screening in men should start at younger ages than in women and that age should not be the sole criteria for indicating CRC screening. Given that prevention of CRC through the removal of adenomas is very important on CRC screening, the association of our GRS model with the presence of adenomas may have important implications for risk stratification.

Finally, our study has several strengths and limitations. To our knowledge, very few studies have evaluated the discriminatory power of GRS in the development of colorectal adenomas. Although only five SNPs were selected in our GRS, the best predictive model, which includes sex and age, shows an AUC value (0.66) similar to that reported by several previous studies evaluating more SNPs [5–9, 18, 19]. The remaining lifestyle information considered in our study (tobacco and NSAIDS/ASA consumption) did not show a significant improvement in discriminatory accuracy of the model. Age and sex are objective data but smoking and drug consumption are more difficult to ascertain and may be prone to bias. In addition, it is important to point out that in our study, individuals with and without adenomas were matched by family history of CRC [10], and therefore, the relevance of family history of CRC as a variable in the prediction models could not be assessed. In this case, GRS could be applied to individuals with an average CRC risk and first-degree relatives of patients with non-hereditary CRC. Further studies with larger sample sizes in both types of populations and in different geographic areas and ethnic groups are needed in order to validate the relevance of our predictive model on the risk of colorectal adenomas. Nevertheless, we believe that combined risk scores (genetic and environmental) might be very useful tools to define when begin (starting age) and what procedure to use (fecal immunochemical test or colonoscopy) in CRC screening programs.

In summary, we propose a GRS associated with colorectal adenomas based on five SNPs (rs10505477, rs11255841, rs13181, rs4779584, and rs8180040). Combining sex and age to the GRS prediction model significantly increases discriminatory accuracy. Moreover, the GRS shows a different behavior depending on sex and age. Our results point out to an association between higher GRS values in men and individuals older than 50 years. The inclusion of other lifestyle factors such as alcohol consumption, body mass index, or dietary factors might improve the predictive capacity of the models. Moreover, advances and decreasing cost for genetic test might facilitate the use of combined risk

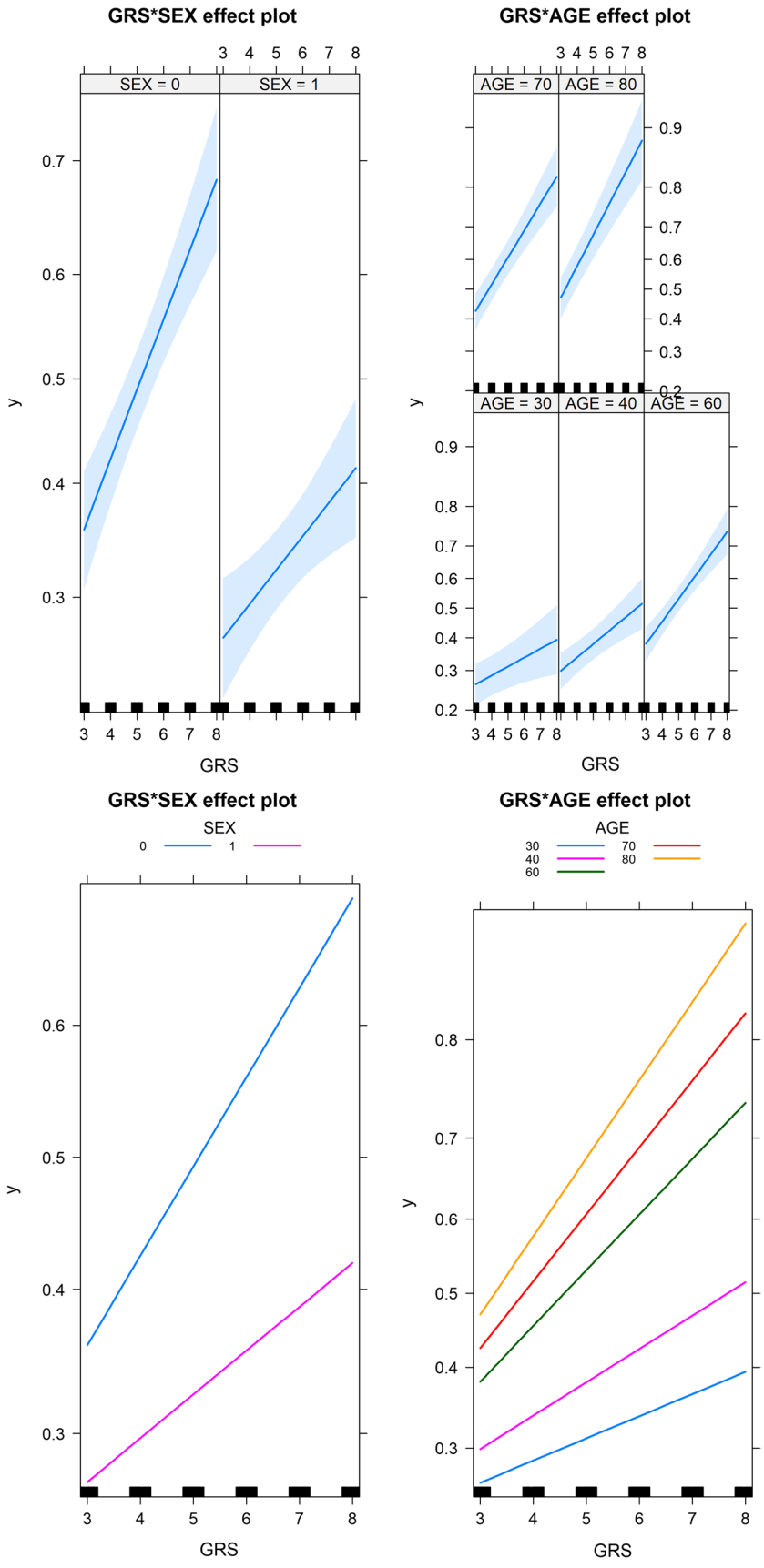


Fig. 4 Visual representation of the interaction between the effects of unweighted GRS and age and sex. In the variable SEX value 0 represents male and value 1 represents female

Table 4 Association between adenomas risk and combination of unweighted GRS and age/sex

GRS values	Sex		Age	
	Female	Male	Age < = 50	Age > 50
	OR (95% CI) *p* value	OR (95% CI) *p* value	OR (95% CI) *p* value	OR (95% CI) *p* value
< = 3	**Ref**	1.800 (0.996–3.936) 0.105	**Ref**	2.182 (1.082–4.629) 0.076
4	0.998 (0.575–1.754) 0.994	2.218 (1.272–3.96) **0.020**	1.993 (0.915–4.505) 0.152	2.165 (1.108–4.477) 0.067
5	0.807 (0.478–1.385) 0.506	2.848 (1.712–4.842) **< 0.001**	1.359 (0.650–2.932) 0.496	2.445 (1.279–4.965) 0.029
6	1.306 (0.784–2.220) 0.397	3.700 (2.222–6.300) **< 0.001**	2.433 (1.215–5.136) **0.041**	3.176 (1.659–6.457) **0.005**
7	2.109 (1.226–3.693) **0.026**	2.953 (1.718–5.176) **0.001**	2.645 (1.279–5.741) **0.032**	3.633 (1.855–7.535) **0.002**
> = 8	3.122 (1.685–5.889) **0.003**	4.580 (2.434–8.808) **< 0.001**	2.675 (1.164–6.364) **0.056**	6.740 (3.253–14.719) **< 0.001**
	Adjusted by age	Adjusted by age	Adjusted by sex	Adjusted by sex

CI confidence interval; *GRS* genetic risk score; *OR* odds ratio; *Ref* reference

prediction models for risk stratification approaches of large populations in a near future.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10620-021-07218-5.

Acknowledgments The authors thank Samantha Arechavaleta for her technical assistance.

Funding This work was supported by Aragonese Society of Digestive Pathology; Spanish Gastroenterology Association (AEG); Aragón Health Research Institute (IIS Aragón); Centro de Investigación biomédica en red en el área enfermedades hepáticas y digestivas (CIBERehd) and Institute of Health Carlos III (ISCIII). The genotyping service was carried out at CEGEN-PRB3-ISCIII; it is supported by grant PT17/0019, of the PE I + D + i 2013–2016, funded by ISCIII and ERDF. This work has been partially funded by the European Social Fund (ESF) and by the European Regional Development Fund (ERDF).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Bray F, Ferlay J, Soerjomataram I et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2018;68:394–424.
2. Huyghe JR, Bien SA, Harrison TA et al. Discovery of common and rare genetic risk variants for colorectal cancer. *NatGenet.* 2019;51:76–87.
3. Schmit SL, Edlund CK, Schumacher FR et al. Novel common genetic susceptibility loci for colorectal cancer. *J Natl Cancer Inst.* 2019;1:146–157.
4. Law PJ, Timofeeva M, Fernandez-Rozadilla C et al. Association analyses identify 31 new risk loci for colorectal cancer susceptibility. *NatCommun.* 2019;14:2154.
5. Jeon J, Du M, Schoen RE et al. Determining risk of colorectal cancer and starting age of screening based on lifestyle, environmental, and genetic factors. *Gastroenterology.* 2018;154:2152-2164.e19.
6. Weigl K, Thomsen H, Balavarca Y et al. Genetic risk score is associated with prevalence of advanced neoplasms in a colorectal cancer screening population. *Gastroenterology.* 2018;155:88-98.e10.
7. Weigl K, Chang-Claude J, Knebel P et al. Strongly enhanced colorectal cancer risk stratification by combining family history and genetic risk score. *ClinEpidemiol.* 2018;10:143–152.
8. Hsu L, Jeon J, Brenner H et al. A model to determine colorectal cancer risk using common genetic susceptibility loci. *Gastroenterology.* 2015;148:1330–9.e14.
9. Ibáñez-Sanz G, Díez-Villanueva A, Alonso MH et al. Risk model for colorectal cancer in Spanish population using environmental and genetic factors: results from the MCC-Spain study. *Sci Rep.* 2017;24:43263.
10. Gargallo CJ, Lanas Á, Carrera-Lasfuentes P et al. Genetic susceptibility in the development of colorectal adenomas according to family history of colorectal cancer. *Int J Cancer.* 2019;144:489–502.
11. Pepe MS, Longton G, Janes H. Estimation and comparison of receiver operating characteristic curves. *Stata J.* 2009;9:1.
12. R Core Team. R. A languaje and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. 2019. Available from: http://www.R-project.org.
13. Kundu S, Aulchenko YS, van Duijn CM, Janssens ACJ. PredictABEL: an R package for the assessment of risk prediction models. *Eur J Epidemiol.* 2011;26:261.
14. Cubiella J, Marzo-Castillejo M, Mascort-Roca JJ et al. Clinical practice guideline. Diagnosis and prevention of colorectal cancer. 20187 Update. *GastroenterolHepatol.* 2018;41:585–596.
15. Rex DK, Boland CR, Dominitz JA et al. Colorectal cancer screening: recommendations for physicians and patients from the U.S. multi-society task force on colorectal cancer. *Am J Gastroenterol.* 2017;112:1016–1030.

16. Abulí A, Castells A, Bujanda L et al. PROCOLON research-group. Genetic variants associated with colorectal adenoma susceptibility. *PLoS One.* 2016;11:e0153084.
17. Dunlop MG, Tenesa A, Farrington SM et al. Cumulative impact of common genetic variants and other risk factors on colorectal cancer risk in 42 103 individuals. *Gut.* 2013;62:871–881.
18. Yarnall JM, Crouch DJ, Lewis CM. Incorporating non-genetic risk factors and behavioural modifications into risk prediction models for colorectal cancer. *Cancer Epidemiol.* 2013;37:324–329.
19. Balavarca Y, Weigl K, Thomsen H, Brenner H. Performance of individual and joint risk stratification by an environmental risk score and a genetic risk score in a colorectal cancer screening setting. *Int J Cancer.* 2020;146:627–634.
20. Zhang B, Shrubsole MJ, Li G et al. Association of genetic variants for colorectal cancer differs by subtypes of polyps in the colorectum. *Carcinogenesis.* 2012;33:2417–2423.
21. Nguyen SP, Bent S, Chen YH, Terdiman JP. Gender as a risk factor for advanced neoplasia and colorectal cancer: a systematic review and meta-analysis. *Clin Gastroenterol Hepatol.* 2009;7:676-81.e1–3.
22. Kolligs FT, Crispin A, Munte A et al. Risk of advanced colorectal neoplasia according to age and gender. *PLoS One.* 2011;6:e20076.
23. Ferlitsch M, Reinhart K, Pramhas S et al. Sex-specific prevalence of adenomas, advanced adenomas, and colorectal cancer in individuals undergoing screening colonoscopy. *JAMA.* 2011;306:1352–1358.

7.4. Estimación del punto de corte óptimo en programas de cribado de cáncer colorrectal

La pandemia de COVID-19 ha supuesto un gran desafío para el sistema de salud pública, ya que obligó a ralentizar o paralizar muchos procedimientos, como los programas de detección de cáncer colorrectal. Esto ha creado una situación posterior en la que hay pacientes que no han sido evaluados a tiempo, además de nuevos pacientes que se suman a la población diana del programa de cribado. Sin embargo, puede existir una oferta insuficiente de unidades endoscópicas para satisfacer las necesidades de la población que necesita ser evaluada. El propósito del estudio [3] que se presenta en esta sección es abordar esta necesidad surgida debido a la pandemia, específicamente, mejorar la estrategia del programa de cribado de cáncer colorrectal en Aragón.

El programa de cribado de cáncer colorrectal consiste en invitar a personas asintomáticas de riesgo medio a realizar una prueba de hemoglobina fecal (FIT). El umbral establecido para esta prueba es de 20 μg de hemoglobina/g de heces, estándar utilizado en España y otros países europeos. En caso de superar ese umbral, se considera la prueba como positiva y se invita al paciente a someterse a una colonoscopia para detectar posibles lesiones. Por tanto, los factores que determinan el nivel de actividad del programa de cribado colorrectal son el número de colonoscopias disponibles, el número de personas invitadas y los resultados del FIT. En concreto, el punto de corte del FIT es el factor clave, puesto que determina quién se somete a colonoscopia y está vinculado con los otros factores. Aumentar el umbral del FIT (20 μg) disminuiría el número de colonoscopias a realizar pero podría implicar un riesgo de pérdida de detección de lesiones de alto riesgo o cánceres. Por otro lado, mantener los valores de corte actuales, con una disponibilidad insuficiente de colonoscopias, retrasaría la detección de lesiones, aumentando el riesgo de detectarlas en etapas más avanzadas.

En nuestro estudio se plantearon estos dos escenarios: modificar o mantener el punto de corte del FIT, y se analizaron sus consecuencias con el objetivo de optimizar la estrategia de cribado de cáncer colorrectal, considerando la situación actual y buscando la opción menos perjudicial. En este sentido, el análisis de diferentes puntos de corte de FIT fue esencial. Como criterio de selección del punto de corte se seleccionó el número de lesiones perdidas (riesgo bajo, medio, alto y cáncer), con el objetivo de minimizar el número esperado de lesiones perdidas y diagnósticos retrasados. Los resultados de nuestro estudio demostraron que, asumiendo un desajuste entre la disponibilidad anual de colonoscopias y la demanda real en la población objetivo, la opción menos perniciosa sería aumentar el punto de corte. Asimismo, nuestros resultados sugirieron considerar en el proceso de selección factores de riesgo adicionales, como el sexo y la edad.

En conclusión, este trabajo plantea un desafío real con consecuencias significativas en el sistema de salud, en el cual proponemos una metodología basada en análisis de puntos de corte. El análisis de puntos de corte según diferentes criterios resulta crucial para la posterior práctica clínica. A continuación, se presenta el artículo publicado derivado de este estudio.

frontiers in Medicine

ORIGINAL RESEARCH
published: 27 July 2021
doi: 10.3389/fmed.2021.712040

Evidence-Based Selection on the Appropriate FIT Cut-Off Point in CRC Screening Programs in the COVID Pandemic

Rocío Aznar-Gimeno[1], *Patricia Carrera-Lasfuentes*[2,3]*, *Rafael del-Hoyo-Alonso*[1], *Manuel Doblaré*[3,4,5,6†] and *Ángel Lanas*[2,3,7,8†]

[1] Department of Big Data and Cognitive Systems, Instituto Tecnológico de Aragón, ITAINNOVA, Zaragoza, Spain, [2] Biomedical Research Networking Center in Hepatic and Digestive Diseases (CIBERehd), Madrid, Spain, [3] Aragón Health Research Institute (IIS Aragón), Zaragoza, Spain, [4] Aragón Institute of Engineering Research (I3A), Zaragoza, Spain, [5] Department of Mechanical Engineering, University of Zaragoza, Zaragoza, Spain, [6] Biomedical Research Networking Center in Bioengineering, Biomaterials and Nanomedicine (CIBERbbn), Madrid, Spain, [7] Department of Medicine, Psychiatry and Dermatology, University of Zaragoza, Zaragoza, Spain, [8] Service of Digestive Diseases, University Clinic Hospital, Zaragoza, Spain

OPEN ACCESS

Edited by:
Cristiano Spada,
Fondazione Poliambulanza Istituto Ospedaliero, Italy

Reviewed by:
Grainne Holleran,
Trinity College Dublin, Ireland
Carlo Senore,
Piedmont Reference Center for Epidemiology and Cancer Prevention, Italy

***Correspondence:**
Patricia Carrera-Lasfuentes
pcarreralasfuentes@gmail.com

[†]These authors share senior authorship

Specialty section:
This article was submitted to Gastroenterology, a section of the journal Frontiers in Medicine

Received: 19 May 2021
Accepted: 29 June 2021
Published: 27 July 2021

Citation:
Aznar-Gimeno R, Carrera-Lasfuentes P, del-Hoyo-Alonso R, Doblaré M and Lanas Á (2021) Evidence-Based Selection on the Appropriate FIT Cut-Off Point in CRC Screening Programs in the COVID Pandemic. Front. Med. 8:712040. doi: 10.3389/fmed.2021.712040

Background: The COVID pandemic has forced the closure of many colorectal cancer (CRC) screening programs. Resuming these programs is a priority, but fewer colonoscopies may be available. We developed an evidence-based tool for decision-making in CRC screening programs, based on a fecal hemoglobin immunological test (FIT), to optimize the strategy for screening a population for CRC.

Methods: We retrospectively analyzed data collected at a regional CRC screening program between February/2014 and November/2018. We investigated two different scenarios: not modifying vs. modifying the FIT cut-off value. We estimated program outcomes in the two scenarios by evaluating the numbers of cancers and adenomas missed or not diagnosed in due time (delayed).

Results: The current FIT cut-off (20-μg hemoglobin/g feces) led to 6,606 colonoscopies per 100,000 people invited annually. Without modifying this FIT cut-off value, when the optimal number of individuals invited for colonoscopies was reduced by 10–40%, a high number of CRCs and high-risk adenomas (34–135 and 73–288/100.000-people invited, respectively) will be undetected every year. When the FIT cut-off value was increased to where the colonoscopy demand matched the colonoscopy availability, the number of missed lesions per year was remarkably reduced (9–36 and 29–145/100.000 people, respectively). Moreover, the unmodified FIT scenario outcome was improved by prioritizing the selection process based on sex (males) and age, rather than randomly reducing the number invited.

Conclusions: Assuming a mismatch between the availability and demand for annual colonoscopies, increasing the FIT cut-off point was more effective than randomly reducing the number of people invited. Using specific risk factors to prioritize access to colonoscopies should be also considered.

Keywords: colorectal cancer, screening fecal-immunological test, decision-making, colonoscopy, adenomas

INTRODUCTION

The COVID pandemic has forced the closure of many colorectal cancer (CRC) screening programs around the world. Most endoscopy units have limited their activity to urgent procedures or to patients with a high suspicion of gastrointestinal cancer (1). Furthermore, resuming normal endoscopic activity has been slow and the number of procedures per room/day has been reduced. This situation poses a challenge for CRC screening programs, because in most cases, endoscopic units will not be able to resume the same activity levels practiced before the pandemic (2, 3).

The key factors that determine the level of activity in CRC screening programs are the number of endoscopic procedures available in the endoscopic units, the number of people invited, and the type of test used. When a fecal occult blood test is used, the cut-off point determines which patients will undergo colonoscopy. Thus, the cut-off point is the most important factor for selecting which people are invited, and it is directly related to the other factors. It is difficult to determine the cut-off point, because raising this value increases the risk of missing a number of high-risk lesions or cancers (4). On the other hand, maintaining the current cut-off values with insufficient colonoscopy availability will delay the inclusion of patients in screening programs, due to the current restrictions, which will cause an undetermined delay in the diagnosis of high-risk lesions and cancer (5).

In this study, we analyzed, from a temporal perspective, the CRC screening program outcomes for the Aragón region (Spain). We aimed to design a general methodology and provide data that might facilitate evidence-based decisions by managers and health authorities on how to reinitiate CRC screening. Our approach was to optimize the strategy for screening a population for CRC, with the objective of minimizing the expected number of missed lesions and delayed diagnoses.

MATERIALS AND METHODS

Study Population

We evaluated data on individuals that participated in the CRC screening program in the Aragon region (Spain) between February 2014 (the start of the population-level program) and November 2018. Individuals invited to the program were at medium risk, aged 60–70 years, had no family history of CRC, had no previous colonoscopy in the previous 5 years, and had no known colonic diseases, colectomy, or irreversible terminal diseases. The screening program was originally planned (before the pandemic) to extend to patients aged 50–59 years, within the universal health system. However, those individuals were excluded from the first round of invitations to maximize the benefits of the program, because endoscopic units were already busy coping with symptomatic patients.

Fecal Immunochemical Test, Colonoscopy, and Lesions

Invited individuals that agreed to participate in the program had to be asymptomatic. They underwent selection, based on a fecal immunochemical test (FIT) (FOB Gold®; SENTiFIT; Sysmex-Sentinel CH. SpA, Barcelona,Spain). The cut-off value used during the program was 20 μg hemoglobin/g feces, which is the standard used in Spain and most European countries (6).

Patients with a negative FIT result were temporarily excluded from the program for 2 years. Otherwise, the patient was invited to undergo a colonoscopy. When the colonoscopy did not show any lesion, the patient was temporarily excluded from the program for 10 years. However, when either adenomas or cancers were detected, the patient was permanently excluded from the program and transferred to either the gastrointestinal outpatients or ward for follow-up or treatment.

Colonoscopy, Histologic Examination, and Definitions

Colonoscopies were performed by experienced gastroenterologists from different units of Digestive Diseases Services in the community. The quality standards were established by the European Society of Gastrointestinal Endoscopy (7, 8). Any polypoid lesion detected in the procedure was removed and classified by an experienced pathologist. The classes were established by the Spanish Network of Cancer Screening Programs (http://www.cribadocancer.es/), based on the European guidelines for quality assurance in CRC screening and diagnosis (9).

The study end points were the lesions, which were classified according to the following grades:

1. "Low-risk adenomas," defined as 1–2 tubular adenomas <1 cm with low grade dysplasia.
2. "Intermediate-risk adenomas," defined as ≥ 3 adenomas, or adenomas ≥ 1 cm, with a villous histology or high-grade dysplasia.
3. "High-risk adenomas," defined as ≥ 10 adenomas or adenomas ≥ 2 cm (mutually exclusive with intermediate-risk adenoma > 1 cm)
4. "Colorectal cancer," defined as any invasive cancer of the colorectal mucosa that reached the submucosa, regardless of the subsequent stage in the TNM classification[1].

Data Curation

We collected data on program participants from data recorded in the program database by different staff members and data uploaded from external databases. Endoscopists of the different centers recorded the data on endoscopic findings. The medical staff in the CRC program recorded data related to the patients, the histological analysis of specimens obtained from endoscopy or surgery, and the treatment given after cancer detection. Therefore, our first step was to perform an extensive evaluation of

[1]Available online at: https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/staged.html (accessed: Januray 29, 2021).

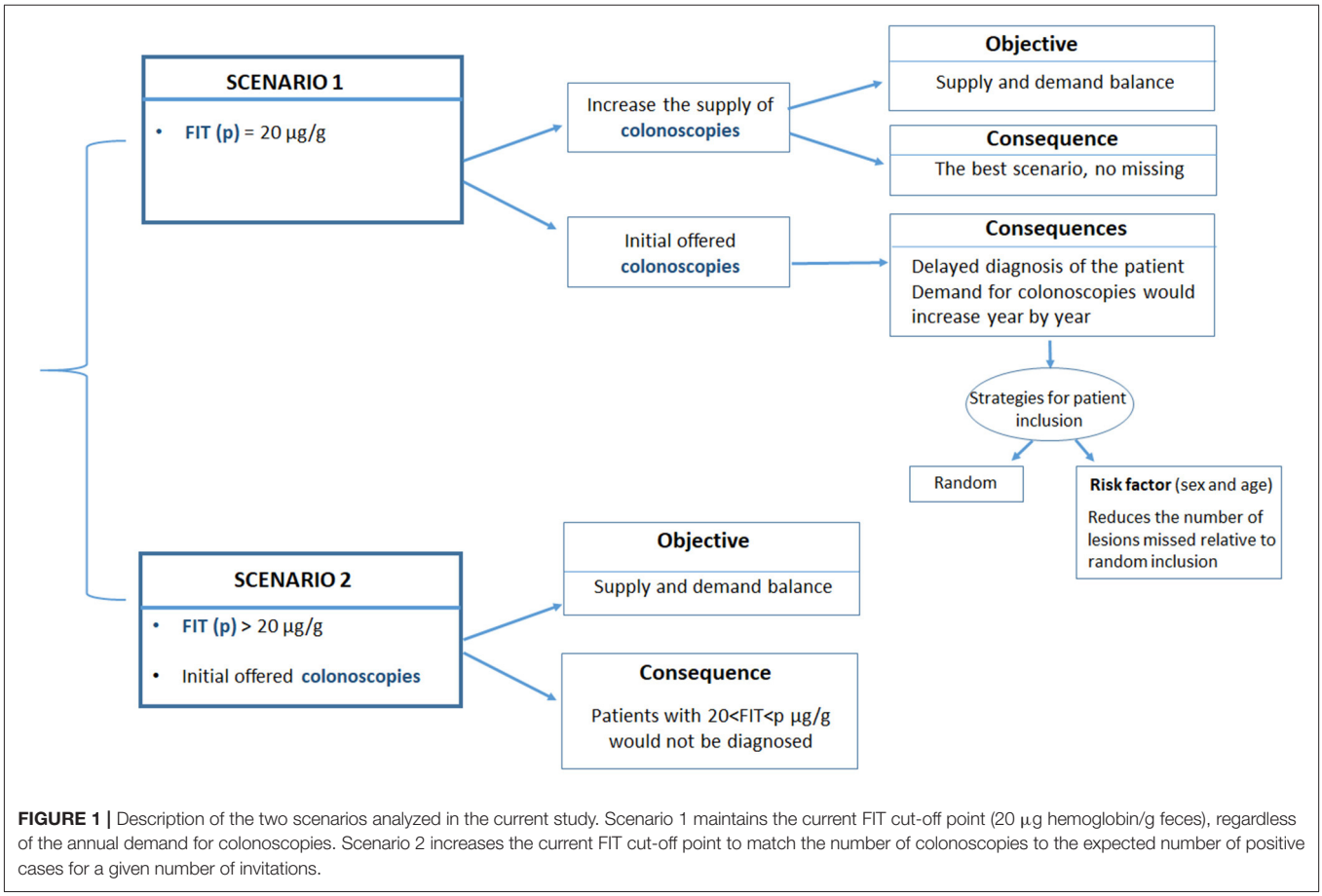


FIGURE 1 | Description of the two scenarios analyzed in the current study. Scenario 1 maintains the current FIT cut-off point (20 µg hemoglobin/g feces), regardless of the annual demand for colonoscopies. Scenario 2 increases the current FIT cut-off point to match the number of colonoscopies to the expected number of positive cases for a given number of invitations.

the quality of the information stored and to carry out a curation process to procure a clean source of information. Detailed information on this process can be found elsewhere (10). Once the data curation was performed, the resulting database could be used for analyses of any particular subpopulation, defined by sex, age, medical district, province, etc.

The methodology used to aid decision-making was based on the relationship between the number of people invited to the program, the number of annual colonoscopies available, and the established FIT cut-off point. Each of these values could be estimated, based on the other two values. Additionally, we analyzed data on the population screened between 2014 and 2018 to estimate the expected number of lesions missed or diagnoses delayed, based on the projected parameters of the three main variables (number of people invited, number of annual colonoscopies performed, and the established cut-off point).

Hypotheses and Analysis

In this study, we assumed that the behavior of the entire target population was homogeneous over time, which included the following corollaries:

1. The number of tests performed (with respect to the number of patients invited, i.e., rate of participation) remained constant over time.
2. The distribution of the stool blood concentration in the population or subpopulation did not change with time; accordingly, the rate of positive tests with respect to the population invited also remained constant over time.
3. The percentage of lesions and of the relative distribution of the risk characteristics (i.e., low risk, medium risk, high risk, and cancer) in the participating population did not change over time.

Scenarios

We analyzed two potential scenarios (**Figure 1**):

1. The current cut-off point was 20 µg hemoglobin/g feces, regardless of the annual demand for colonoscopies.
2. The current cut-off point was increased to match the number of individuals invited for colonoscopies; here, the demand was derived from the expected number of positive tests for a given number of invitations.

Scenario 1 was optimal, providing that sufficient colonoscopies were available to meet the yearly demand. When the demand could not be met, a certain percentage of individuals with potential neoplastic lesions will not be diagnosed, because they would not be invited to undergo a colonoscopy in the corresponding year. The estimated number of patients in need and the types of lesions that were not diagnosed was derived

TABLE 1 | Screening results for the study population and subpopulations.

	Total	Male	Female	Age [60–65] y	Age [65–70] y
Invited (*n*)	**146,811**	71,363	75,448	71,621	75,190
FIT Performed[a]	**76,452 (52.1%)**	36,971 (51.8%)	39,481 (52.3%)	35,476 (49.5%)	40,976 (54.5%)
Positive FIT[b]	**9,699 (6.6%)**	5,748 (8.1%)	3,951 (5.2%)	4,335 (6.1%)	5,364 (7.1%)
Colonoscopies[c]	**9,139 (6.2%)**	5,433 (7.6%)	3,706 (4.9%)	4,051 (5.7%)	5,088 (6.8%)
Neoplastic lesions[d]	**4,823 (52.8%)**	3,315 (61 %)	1,508 (40.7%)	2,126 (52.5%)	2,697 (53%)
Low risk lesions[e]	**1,399 (29%)**	825 (24.9%)	574 (38.1%)	656 (30.9%)	743 (27.5%)
Medium risk lesions[e]	**1,959 (40.6%)**	1,382 (41.7%)	577 (38.3%)	869 (40.9%)	1,090 (40.4%)
High risk lesions[e]	**1,003 (20.8%)**	767 (23.1%)	236 (15.6%)	413 (19.4%)	590 (21.9%)
Cancer lesions[e]	**462 (9.6%)**	341 (10.3%)	121 (8%)	188 (8.8%)	274 (10.2%)

[a,b,c]*Number and percentage calculated based on the total number of invitations.*

[b]*FIT cut-off point >20-μg hemoglobin/g feces.*

[d]*Number of people with lesions and percentage calculated based on the total number of people with lesions.*

[e]*Number of people with lesions and percentage calculated based on the total number of colonoscopies.*

from the data obtained from the cohort of individuals screened between 2014 and 2018.

In Scenario 1, those individuals that were not eventually invited due to an insufficient colonoscopy offer (excess target group) were added to the new cohort invited the next year. Then, and in this scenario, we evaluated different strategies for selecting individuals for screening in the subsequent year. In the first "random strategy," all patients (new invitations and the excess target group of previous year) are randomly selected according to the same criteria (**Figure 1**), that is, all patients assigned to a particular year have the same probability of undergoing a colonoscopy, independent of the year of their first invitation to the screening program. In a second "prioritization strategy," the excess target group from the previous year would be attended first, and then individuals in the new yearly cohort would be attended. An extension to the prioritization strategy was to prioritize access to colonoscopy by considering risk factors other than the FIT, such as age or sex.

In this study, we analyzed the possibility of prioritizing invitations based on sex, because men were demonstrated to have a significantly higher risk of lesions than women. This approach aimed to show the importance of establishing a risk index that included both the stool blood concentration and other risk factors to minimize the number of high-grade lesions missed.

In Scenario 2, the cut-off point (cop) was set, based on the number of colonoscopies to be offered. Then, we estimated the rate and number of lesions that would be missed in that year. Patients with a FIT result below the new "cop" but a FIT result > 20 μg hemoglobin/g feces, will not be identified in this scenario.

The estimations for each of these two scenarios are presented considering 100,000 people invited. However, it would be easy to scale the results to any other number of invitations. After the pandemic, the number of colonoscopies actually performed has been drastically reduced. We considered different potential reductions in the offer of colonoscopies. Specifically, we analyzed 60–90% of the total number of colonoscopies that would have met the demand.

Statistical Analysis

To demonstrate the potential effects of potential risk factors to be considered, we analyzed several subpopulations, defined by sex or age. Particularly, we compared the rate of positive tests in the invited population, the rate of lesion detection in colonoscopies, and the distribution of different risk levels among the lesions detected between different subpopulations. For this comparative analysis, we performed the compare proportions test. The level of significance in the study was set to 0.05. Analyses were performed with the R programming language (The R Foundation for statistical computing, Vienna, Austria) (11).

RESULTS

Outcomes

The screening results are shown in **Table 1** (2014–2018); 6.6% of the invited individuals had a positive test (FIT $\geq$ 20 μg hemoglobin/g feces). Among the individuals that underwent a colonoscopy, 52.8% had detectable lesions, and of these, 30.4% were classified as high-risk or had cancer. The different sex and age groups had different rates of lesion detection. The rates of positive FIT were significantly higher ($p < 0.001$) for men than women and higher for the older, compared to the younger age group. Also, the rates of detected neoplastic lesions were

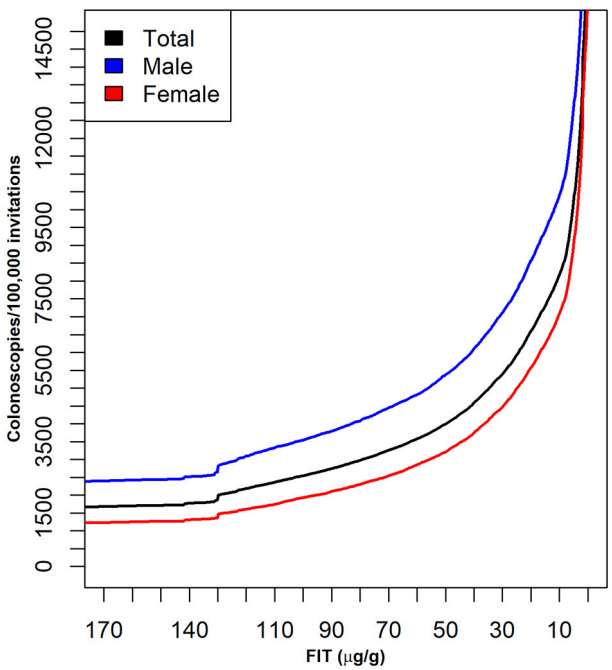


FIGURE 2 | The numbers of colonoscopies required with different FIT cut-off values. Data are shown for the whole population and for sex groups. X axis: FIT (μg hemoglobin/g feces); Y axis: number of patients with a positive test, per 100,000 invitations.

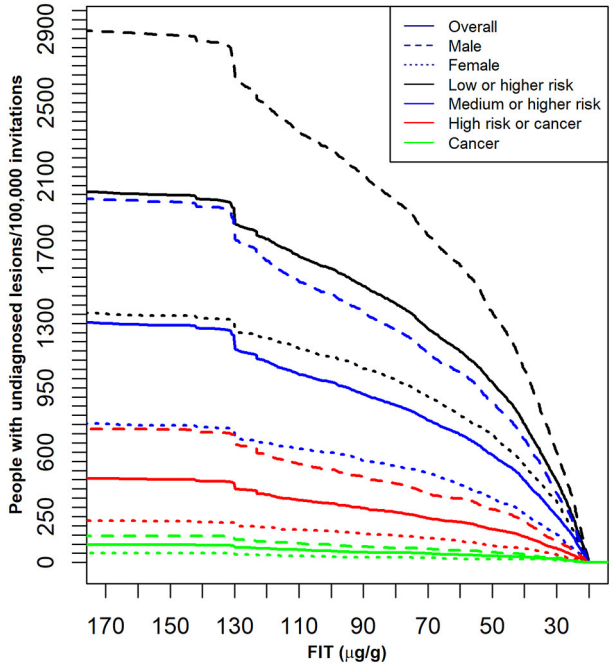


FIGURE 3 | The number of individuals with undiagnosed lesions, based on the indicated FIT cut-off points. Dark lines: the whole population (Overall); Different colors: different lesion risk classes; dashed and dotted lines: different sex groups for each lesion risk class; X axis: FIT: (μg hemoglobin/g feces). Y axis: the number of people with undiagnosed lesions per 100,000 invitations.

significantly different ($p < 0.001$) between men and women (0.61 vs. 0.40). The distribution of lesion types was also significantly different ($p < 0.001$) among the different combinations of sex and age (**Table 1**) The stool blood concentrations showed small differences over the different years (2014–2018), and we expected the curves to converge to a steady state level. Therefore, we assumed that those differences will not impact the qualitative trends, or the conclusions drawn from the analysis reported here. The number of colonoscopies performed was essentially linearly correlated with the population invited, assuming no important variations in the percentage of people that accepted the invitation, underwent the FIT, and had a positive result.

Figure 2 shows the curves used to determine a cut-off point for matching a given availability of colonoscopies per 100,000 invitations. **Figure 3** shows the number of people in each lesion risk class with undiagnosed lesions for each cut-off point considered Sex, as expected, was remarkably discriminating. With the same cut-off point, the number of men with undiagnosed high-risk lesions or cancer was very close to the number of women with medium- or higher-risk lesions. Age was also discriminating, but to a lesser extent (not shown).

Scenario 1

With a constant cut-off point of 20 μg of hemoglobin/g feces, 6,606 colonoscopies would be required for each 100,000 people invited annually. However, if that number of colonoscopies was not available, and the cut-off point remained at 20 μg hemoglobin/g feces, the number of undiagnosed patients with positive tests would accumulate over the subsequent years, and individuals would not be invited in due time. We evaluated the following proportions of colonoscopy availability: 90–60% where 100% was 6,606 colonoscopies/100,000 people invited and required for the analyzed population.

Table 2 shows a comparison of the undiagnosed or delayed risk lesions for different colonoscopy availabilities in each scenario, and for each call criterion for the population of 1 year. We have explored an offer of colonoscopies of 90, 80, 70, and 60% of the 6,606 colonoscopies required for each 100,000 people invited annually when assuming the cut-off point of equilibrium (20 μg hemoglobin/g feces).

Considering a period of 5 years, and assuming 100,000 invitations each year and the same colonoscopy availability in the whole 5 years period, **Figure 4** shows the estimated number of patients with different types of colorectal lesions, whose diagnosis would be delayed by at least 1, 2, 3, 4, or 5 years, given the different scenarios of colonoscopy reduction. The graphs show that the lower the colonoscopy capacity the higher the number of lesions whose diagnosis will be delay for at least 1–5 years. For example, with 90% of equilibrium colonoscopy availability, no

TABLE 2 | Undiagnosed lesions per year considering scenarios 1 and 2 and the two call criteria in scenario 1.

		Lesion risk class *N* (%)[a]				
% Of available colonoscopies	**Scenarios**	**Low-risk**	**Medium-risk**	**High-risk**	**Cancer**	**Total lesions**
90%	**Scenario 1** Random call criterion	101 (10%)	141 (10%)	73 (10%)	34 (10%)	349 (10%)
	Scenario 1 Prioritized call criterion	102 (10.13%)	103 (7.29%)	42 (5.82%)	22 (6.47%)	269 (7.71%)
	Scenario 2	97 (9.59%)	98 (6.93%)	29 (3.99%)	9 (2.67%)	233 (6.68%)
80%	**Scenario 1** Random call criterion	202 (20%)	283 (20%)	145 (20%)	67 (20%)	697 (20%)
	Scenario 1 Prioritized call criterion	205 (20.25%)	206 (14.57%)	85 (11.63%)	44 (12.94%)	540 (15.48%)
	Scenario 2	203 (20.08%)	198 (14%)	62 (8.53%)	21 (6.23%)	484 (13.87%)
70%	**Scenario 1** Random call criterion	303 (30%)	424 (30%)	218 (30%)	101 (30%)	1,046 (30%)
	Scenario 1 Prioritized call criterion	307 (30.39%)	309 (21.87%)	127 (17.45%)	65 (19.41%)	808 (23.16%)
	Scenario 2	298 (29.48%)	300 (21.22%)	108 (14.86%)	29 (8.6%)	735 (21.07%)
60%	**Scenario 1** Random call criterion	404 (40%)	568 (40%)	288 (40%)	135 (40%)	1,395 (40%)
	Scenario 1 Prioritized call criterion	410 (40.51%)	412 (29.15%)	169 (23.26%)	87 (25.87%)	1,078 (30.9%)
	Scenario 2	398 (39.37%)	410 (29%)	145 (19.94%)	36 (10.68%)	989 (28.35%)

[a] *Percentage of non-diagnosed lesions per year with respect to the total number of estimated lesions of the same type in 100,000 invitations/year for the 2 scenarios and 2 call criteria considered.*

lesions will be delayed for more than 3 years (**Figure 4**). However, the risk of the lesions may progress with time, so the proportions of lesions with higher risk might increase with time, although this issue has not been addressed in this study.

We found different results, depending on the selection strategy for re-inviting patients that could not undergo the timely colonoscopy (the excess target group). We compared two strategies; one employed the random call criterion, where the patients were invited at random (**Figure 4**). The second strategy employed the prioritized call criterion, which prioritized men for the invitations (**Figure 5**). We conducted this exercise to understand the effect of a potential risk factor, and we selected sex as a risk factor without regard to any possible policy decisions.

Scenario 2

We found that the current cut-off point of 20 μg hemoglobin/g feces would be acceptable for 6,606 colonoscopies per 100,000 people invited annually, when there was no restriction on availability. However, this cut-off point would not meet the demand, when the colonoscopy availability was reduced to 90, 80, 70, or 60% of the demand. In those cases, to meet the demand, the FIT cut-off points would have to be raised to 25, 32, 40, and 51 μg hemoglobin/g feces, respectively (**Figure 2**). However, raising the cut-off point could result in missed diagnoses.

The estimated number of people with neoplastic lesions that would not be diagnosed when considering these cut-off points were extracted from **Figure 3** and are shown in **Table 2**. For example, when 25 μg hemoglobin/g feces was the FIT cut-off point, 38 high-risk lesions or cancers and 9 cancers would not be diagnosed annually. These numbers represented, respectively, 3.6 and 2.7% of the lesions identified with the 20 μg hemoglobin/g feces cut-off in those risk classes.

DISCUSSION

The COVID-19 pandemic stopped or slowed many CRC screening programs. This interruption will have a great impact on the number of CRCs diagnosed and on the prognosis of newly diagnosed cases, which will, presumably, be diagnosed at an advanced stage (12–14). Resuming these programs is becoming a priority, but unfortunately, the new rules for preventing COVID-19 transmission have reduced the availability of colonoscopies in most health system (1, 2, 15). This situation has complicated the chronic problem, present before the pandemic, of insufficient colonoscopy availability in many public health systems. Indeed, waiting lists for endoscopies (for patients with symptoms) and CRC screening have been the norm (16–18). Here, we evaluated two scenarios and options for coping with this problem in CRC screening programs that use FIT to select patients for diagnostic colonoscopies.

We proposed a simple and flexible methodology that can be extended to other sub-populations and even to other screening programs for decision-making, which becomes important in this or other situations.

Previous studies have evaluated the impact of raising the FIT cut-off point to match the demand to the availability of

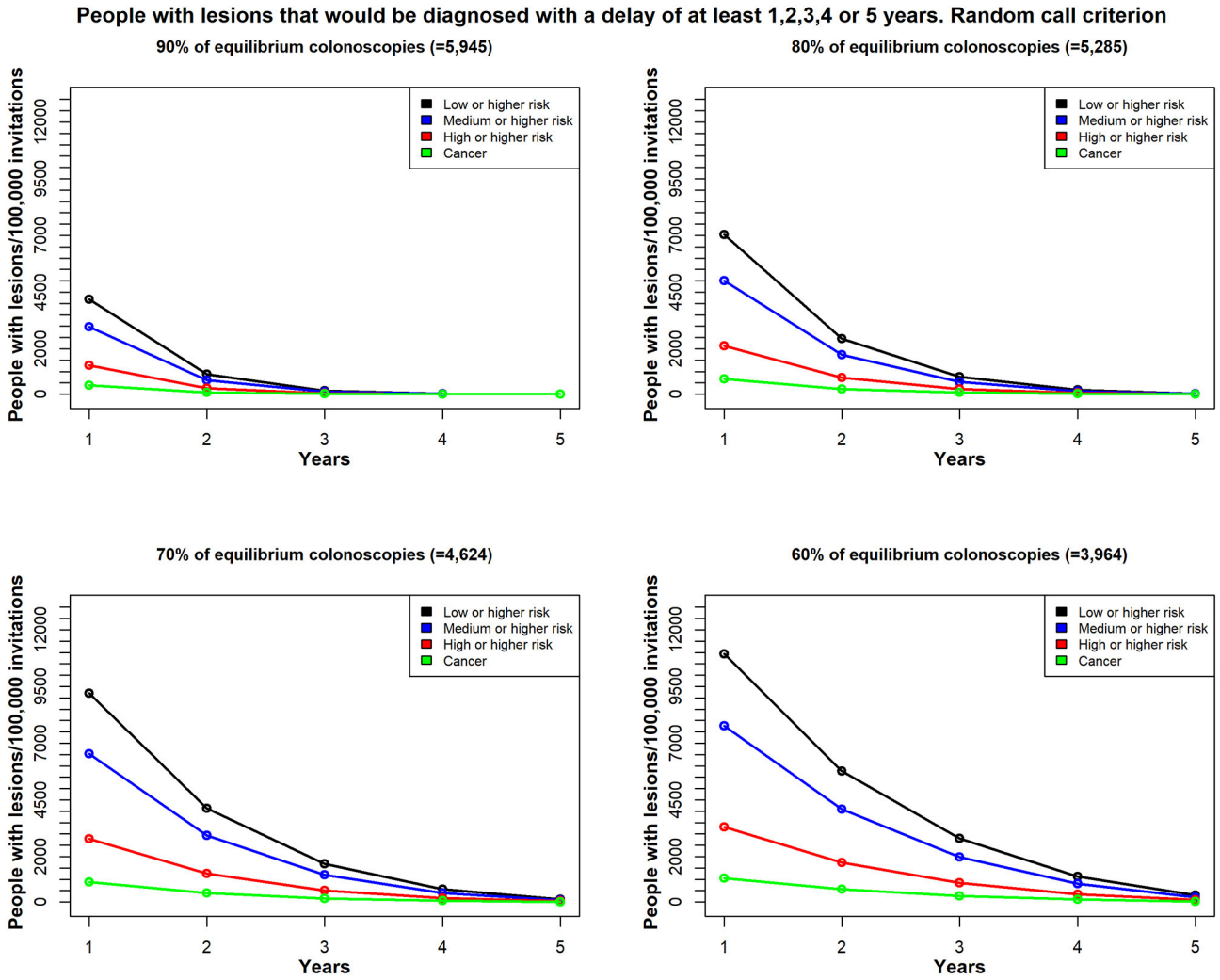


FIGURE 4 | In Scenario 1, the time delay, due to unavailable colonoscopies, for a diagnosis among patients with colorectal lesions in each risk class; patients were re-invited, based on the random call criterion. Total demand for colonoscopies = 6,606/100,000 individuals per year; estimates reflect (*top left*) 90% availability (n = 5,945); (*top right*) 80% availability (n = 5,285); (*bottom left*) 70% availability (n = 4,624); and (*bottom right*) 60% availability (n = 3,964); Y-axis values reflect the estimated number of patients with lesions that would be diagnosed with a delay of at least 1 year, 2 years up to 5 years; X-axis values reflect the minimum delay (years).

colonoscopies in CRC screening programs (4, 19). However, this option is not always followed in countries that use FIT as the screening test, because it implies that lesions with insufficient bleeding might not test positive, and thus, would be missed. To address that concern, in the present study, we evaluated two alternative scenarios for reducing the number of patients screened, and we provided a comparison of the proportions, types, and numbers of CRC and adenoma lesions that would be missed in each scenario.

Maintaining the cut-off point when the system cannot provide a sufficient number of colonoscopies leads to delays in diagnoses that could increase exponentially over the years. Thus, a large number of patients with neoplastic lesions would not be treated, until after a significant delay or when they become symptomatic. Delays can lead to lesion progression to cancer or advanced stages, and even death, which goes against the objective of screening (20). However, we found that the poor results of scenario 1, where the cut-off value was not modified, could be reduced by introducing additional factors into the selection process, such as the sex of the target population. We chose sex prioritization, because the proportion of serious and more advanced lesions was higher among men than among women. Age is another significant risk factor, and both these factors could be introduced to optimize the results of the screening program without modifying the cut-off point. Thus, in Scenario 1, fewer target patients were missed with prioritization than with the random criterion. In both cases, the magnitude and impact of the method on the numbers of undiagnosed lesions that accumulated over the years depended on how much the availability of colonoscopies was reduced, compared to the ideal

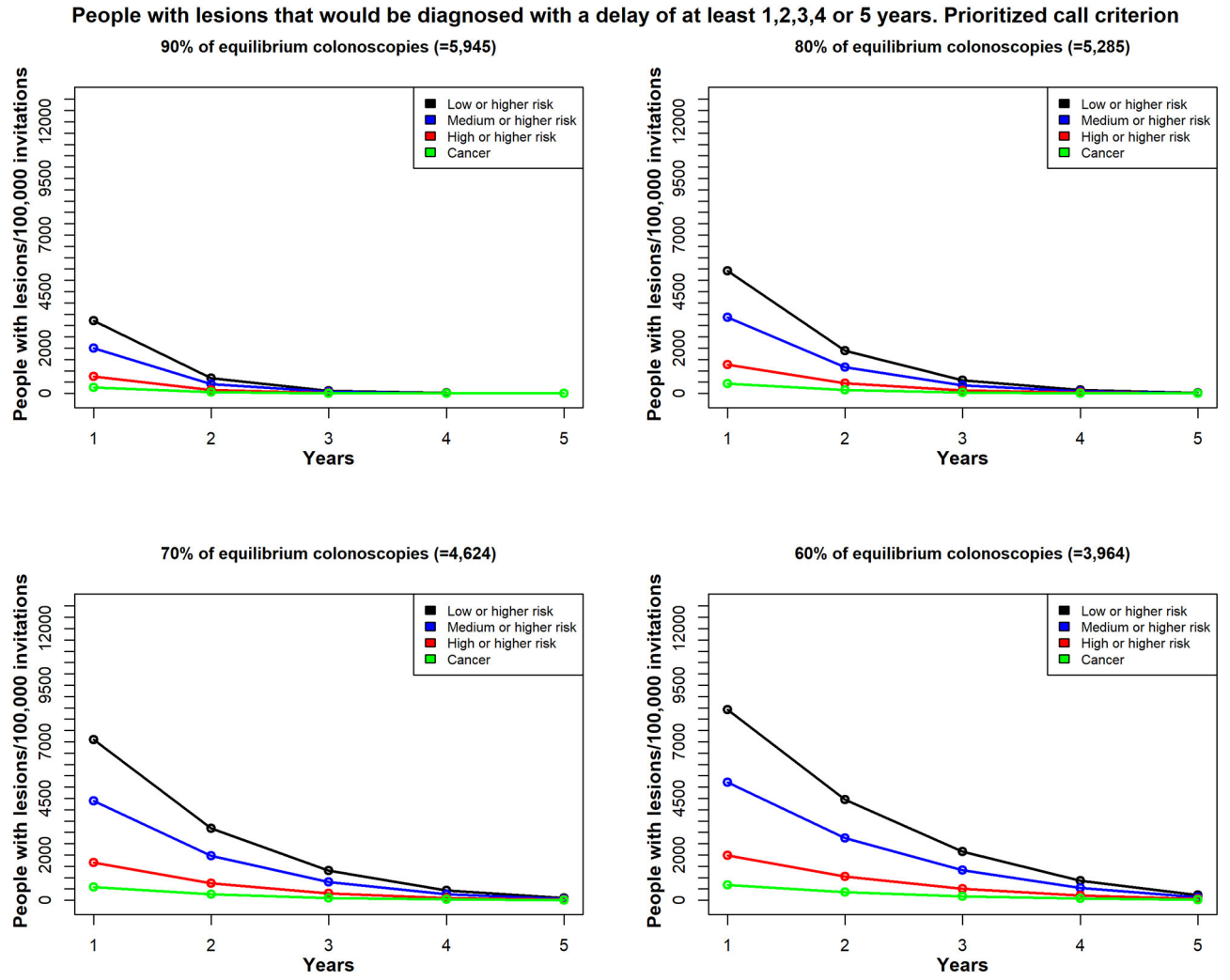


FIGURE 5 | In Scenario 1, the time delay, due to unavailable colonoscopies, for a diagnosis among patients with colorectal lesions in each risk class; patients were re-invited, based on the prioritized call by sex (men) criterion. Total demand for colonoscopies = 6,606/100,000 individuals per year; estimates reflect (*top left*) 90% availability (n = 5,945); (*top right*) 80% availability (n = 5,285); (*bottom left*) 70% availability (n = 4,624); and (*bottom right*) 60% availability (n = 3,964); Y-axis values reflect the estimated number of patients with lesions that would be diagnosed with a delay of at least 1 year, 2 years up to 5 years; X-axis values reflect the minimum delay (years).

number; nevertheless, the random criterion always had a greater impact than the prioritization criterion.

In contrast, raising the FIT cut-off point would not cause an accumulation of delayed colonoscopies. However, a number of lesions would not be diagnosed until the next round(s) or when the lesion becomes symptomatic. The advantage of this approach is that the program will cover the target population every 1–2 years, and no factors need to be introduced "a priori" to optimize the results. In contrast, maintaining the FIT cut-off point would significantly delay the screening of the whole target population. In comparing these strategies, our results showed that increasing the FIT cut-off point should be preferred, because the number of lesions missed was systematically lower than those missed by maintaining the FIT cut-off and delaying the screening.

Our study had some limitations. First, we focused on the 60–70-year-old population, because it was the population screened in our current regional program. This population probably had a higher number of lesions than individuals aged 50–60 years, who are typically included in most CRC screening programs. Nevertheless, we believe that the methodology and the main conclusions were valid; only the number of lesions missed or delayed might change in different populations.

Additionally, we assumed that the proportion of people that accepted the invitation to undergo a colonoscopy, and the proportion of lesions found in colonoscopies would remain stable over time. Clearly, this assumption might not be upheld, but with progressive implementation of the screening program,

the level of participation and the outcomes should stabilize over time.

Increasing the cut-off of FIT will increase the number of interval cancers and the progression of some adenomas to cancer after negative FITs, especially if the increased cut-off level is maintained over time, which implies that our study may underestimate the effects of raising the cut-off. However, it is also true that the risk of detecting a colon cancer increases with the FIT value, and the effect may be higher for low-risk lesions (18). Also, CRC screening programs triggers the demand for surveillance colonoscopy, which will affect the colonoscopy capacity in the subsequent years. This effect has not been taken into consideration in our analysis since it is difficult to make realistic estimation.

In conclusion, we provided an evidence-based methodology that might facilitate decision-making in CRC screening programs, when the demand exceeds the availability of colonoscopies, under any potential circumstances, such as the current situation with the COVID-19 pandemic. Our results showed that, assuming a mismatch between the annual availability of colonoscopies and the actual demand in the target population, it is better to increase the cut-off point than to maintain the cut-off point and call people at random to undergo the colonoscopy. In addition, our findings argued for the inclusion of risk factors, like sex and probably age, in the selection process, to minimize the number of missed high-risk lesions.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: Data sharing is not applicable to this article as no new data were created in this study. Requests to access these datasets should be directed to pcarreralasfuentes@gmail.com.

ETHICS STATEMENT

This study was reviewed and approved by the Regional Ethical Committee of Aragón (CEICA). Written informed consent from the participants was not required since this study is retrospective and used data stored in databases, which were anonymized for data analysis.

AUTHOR CONTRIBUTIONS

ÁL and MD coordinated the whole work, reviewed the drafted manuscript, discussed the results, and approved the submitted manuscript. ÁL identified the clinical problem, defined the associated objectives, revised the clinical results, and got the data. MD supervised the technical approach and the corresponding results. RA-G and PC-L prepared the data and did the statistical analysis. Rd-H-A made the computations and graphics. ÁL, MD, and RA-G drafted the manuscript, while PC-L and Rd-H-A also contributed to the final version. ÁL and Rd-H-A obtained funding. All authors contributed to the article and approved the submitted version.

FUNDING

This work has been partially funded by the European Social Fund (ESF) and by the European Regional Development Fund (ERDF).

ACKNOWLEDGMENTS

This study was made possible thanks to the agreement between the Health Department of the Aragón Government, IIS Aragon and ITAINNOVA to explore the CRC screening program. The authors also thank all participants and staff involved in the program for making this study possible. Finally, the authors thank Mabel Cano for her inestimable collaboration in providing anonymized access to the Regional Database in CRC screening program.

REFERENCES

1. Parasa S, Reddy N, Faigel DO, Repici A, Emura F, Sharma P. Global impact of the COVID-19 pandemic on endoscopy: an international survey of 252 centers from 55 countries. *Gastroenterology*. (2020) 159:1579–81. doi: 10.1053/j.gastro.2020.06.009
2. Gupta S, Shahidi N, Gilroy N, Rex DK, Burgess NG, Bourke MJ. Proposal for the return to routine endoscopy during the COVID-19 pandemic. *Gastrointest Endosc*. (2020) 92:735–42. doi: 10.1016/j.gie.2020.04.050
3. Repici A, Pace F, Gabbiadini R, Colombo M, Hassan C, Dinelli M, et al. Endoscopy units and the coronavirus disease 2019: outbreak: a multicenter experience from Italy. *Gastroenterology*. (2020) 159:363–6.e3. doi: 10.1053/j.gastro.2020.04.003
4. Wieten E, Schreuders EH, Nieuwenburg SA, Hansen BE, Lansdorp-Vogelaar I, Kuipers EJ, et al. Effects of increasing screening age and fecal hemoglobin cutoff concentrations in a colorectal cancer screening program. *Clin Gastroenterol Hepatol*. (2016) 14:1771–7. doi: 10.1016/j.cgh.2016.08.016
5. Zorzi M, Senore C, Turrin A, Mantellini P, Visioli CB, Naldoni C, et al. Italian colorectal cancer screening survey group: appropriateness of endoscopic surveillance recommendations in organized colorectal cancer screening programmes based on the faecal immunochemical test. *Gut*. (2016) 65:1822–8. doi: 10.1136/gutjnl-2015-310139
6. Quintero E, Castells A, Bujanda L, Cubiella J, Salas D, Lanas A, et al. Colonoscopy versus fecal immunochemical testing in colorectal-cancer screening. *N Engl J Med*. (2012) 366:697–706. doi: 10.1056/NEJMoa1108895
7. Rembacken B, Hassan C, Riemann JF, Chilton A, Rutter M, Dumonceau JM, et al. Quality in screening colonoscopy: position statement of the European Society of Gastrointestinal Endoscopy (ESGE). *Endoscopy*. (2012) 44:957–68. doi: 10.1055/s-0032-1325686
8. Mangas-Sanjuan C, Zapater P, Cubiella J, Murcia Ó, Bujanda L, Hernández V, et al. Importance of endoscopist quality metrics for findings at surveillance colonoscopy: the detection-surveillance paradox. *United Eur Gastroenterol J*. (2018) 6:622–9. doi: 10.1177/2050640617745458
9. Atkin WS, Valori R, Kuipers EJ, Hoff G, Senore C, Segnan N, et al. European guidelines for quality assurance in colorectal cancer screening and diagnosis. First Edition–Colonoscopic surveillance following adenoma removal. *Endoscopy*. (2012) 44:SE151–63. doi: 10.1055/s-0032-1309821
10. Aznar-Gimeno R, Carrera-Lasfuentes P, Rodrigalvarez-Chamorro V, Del Hoyo-Alonso R, Lanas A Doblaré M. Analysis and curation of the database

of a colo-rectal cancer screening program. In: S. Kumar, editor. *Data Integrity and Quality*. Rijeka: IntechOpen (2021). doi: 10.5772/intechopen.95899. Available online at: https://www.intechopen.com/books/data-integrity-and-quality/analysis-and-curation-of-the-database-of-a-colo-rectal-cancer-screening-program

11. Team RC. *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2020). Available online at: http://www.r-project.org/index.html (accessed: Jan 25, 2021).
12. Maida M. Screening of gastrointestinal cancers during COVID-19: a new emergency. *Lancet Oncol*. (2020) 21:e338. doi: 10.1016/S1470-2045(20)30341-7
13. Suárez J, Mata E, Guerra A, Jiménez G, Montes M, Arias F, et al. Impact of the COVID-19 pandemic during Spain's state of emergency on the diagnosis of colorectal cancer. *J Surg Oncol*. (2021) 123:32–36. doi: 10.1002/jso.26263
14. Hunt RH, East JE, Lanas A, Malfertheiner P, Satsangi J, Scarpignato C, et al. COVID-19 and Gastrointestinal Disease: implications for the Gastroenterologist. *Dig Dis*. (2020) 9:1–21. doi: 10.1159/000512152
15. Dekker E, Chiu HM, Vogelaar IL. Colorectal cancer screening in the novel coronavirus disease-2019 era. *Gastroenterology*. (2020) 159:1998–2003. doi: 10.1053/j.gastro.2020.09.018
16. Petersen MM, Ferm L, Kleif J, Piper TB, Rømer E, Christensen IJ, et al. Triage may improve selection to colonoscopy and reduce the number of unnecessary colonoscopies. *Cancers*. (2020) 12:2610. doi: 10.3390/cancers12092610
17. Leddin D, Armstrong D, Borgaonkar M, Bridges RJ, Fallone CA, Telford JJ, et al. The 2012 SAGE wait times program: survey of access to gastroenterology in Canada. *Can J Gastroenterol*. (2013) 27:83–9. doi: 10.1155/2013/143018
18. Navarro M, Hijos G, Ramirez T, Omella I, Carrera-Lasfuentes P, Lanas Á. Fecal hemoglobin concentration, a good predictor of risk of advanced colorectal neoplasia in symptomatic and asymptomatic patients. *Front Med*. (2019) 6:91. doi: 10.3389/fmed.2019.00091
19. Kooyker AI, Toes-Zoutendijk E, Opstal-van Winden AWJ, Spaander MCW, Buskermolen M, van Vuuren HJ, et al. The second round of the Dutch colorectal cancer screening program: impact of an increased fecal immunochemical test cut-off level on yield of screening. *Int J Cancer*. (2020) 147:1098–106. doi: 10.1002/ijc.32839
20. Mutneja HR, Bhurwal A, Arora S, Vohra I, Attar BM. A delay in colonoscopy after positive fecal tests leads to higher incidence of colorectal cancer: a systematic review and meta-analysis. *J Gastroenterol Hepatol*. (2021) 36:1479–86. doi: 10.1111/jgh.15381

Conflict of Interest: ÁL was advisors to Sysmex Iberica. Spain.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Capítulo 8

Discusión y Conclusiones

La investigación sobre la estimación y aplicación de modelos de clasificación binaria ha sido un área de interés en los últimos años, con aplicaciones en diversos campos. En concreto, en el ámbito de la biomedicina y la salud, estos modelos son fundamentales para determinar el estado de los pacientes, lo que puede ser crucial para prevenir problemas de salud o establecer estrategias de seguimiento. El objetivo principal es desarrollar modelos con una mayor capacidad predictiva, que se ajusten mejor a los datos y proporcionen predicciones más precisas. No obstante, para que estos modelos sean útiles en la práctica clínica, también es importante considerar su utilidad clínica. En este sentido, la elección y aplicación de un punto de corte óptimo, basado en un objetivo a optimizar, es crucial para convertir la predicción del modelo en decisiones prácticas y útiles. Esto permite clasificar a un paciente como positivo para la enfermedad o condición, facilitando la toma de decisiones clínicas. La investigación presentada en esta memoria ha abordado la estimación de modelos para problemas de clasificación enfocados en el ámbito de la salud y la medicina. Se han propuesto nuevos enfoques y abordado problemas de clasificación en entornos clínicos reales. La evaluación de la capacidad de los modelos fue clave para seleccionar el más adecuado, considerando asimismo la utilidad clínica.

La curva ROC es una herramienta fundamental en la estimación de modelos de clasificación binaria, ampliamente utilizada en el diagnóstico de enfermedades. A partir de esta curva, se derivan medidas resumen que evalúan el rendimiento del modelo, como el AUC y el índice de Youden. En el capítulo 1 de esta tesis, se ha introducido la teoría derivada de la curva ROC, sus propiedades y estimaciones. En el capítulo 2, se han presentado las definiciones y estimaciones de las métricas derivadas (AUC e índice de Youden). Aunque ambas métricas han sido ampliamente utilizadas en estudios clínicos, difieren en su enfoque. Mientras que el AUC evalúa la capacidad discriminatoria general del modelo considerando todos los puntos de corte posibles, el índice de Youden se enfoca en un punto de corte específico cuya optimización maximiza tanto la sensibilidad

como la especificidad del modelo. Por tanto, si el objetivo final es la aplicación del modelo para la ayuda a la toma de decisiones basada en sus predicciones, el índice de Youden proporciona adicionalmente un umbral de decisión específico que permite una clasificación precisa de los pacientes. Esto es crucial en algunas aplicaciones prácticas como el diagnóstico de enfermedades.

En el capítulo 3 se han introducido otros métodos para determinar el punto de corte óptimo, además del índice de Youden. No obstante, cuando no hay un consenso claro sobre la importancia relativa de la sensibilidad y la especificidad, el índice de Youden es un criterio adecuado para seleccionar el punto de corte óptimo, ya que equilibra ambas métricas y sirve también como medida resumen adecuada para evaluar la capacidad predictiva del modelo. En el capítulo 4 se han presentado algoritmos y modelos aparecidos en la literatura para abordar problemas de clasificación. En particular, en la sección 4.1, se describen enfoques desarrollados en la literatura que se basan en la maximización del AUC. Estos enfoques han sido el punto de partida para el desarrollo de otros algoritmos que utilizan otros criterios de optimalidad derivados de la curva ROC (sección 4.2).

Los trabajos principales que forman parte del compendio de esta tesis [4, 8, 7], presentados en los capítulos 5 y 6, tienen como objetivo la estimación de nuevos enfoques para la combinación lineal de biomarcadores continuos en problemas de clasificación binaria bajo optimización del índice de Youden. La motivación que impulsó estas investigaciones fue desarrollar y comparar modelos con mayor capacidad predictiva bajo la maximización del índice de Youden, una métrica que ha recibido menor atención en la literatura pero que es óptima para seleccionar el punto de corte y resumir el rendimiento del modelo en ausencia de consenso.

En este contexto de modelos lineales, autores como Su y Liu [108] propusieron la estimación paramétrica de un modelo lineal que maximiza el AUC bajo la asunción de normalidad. Sin embargo, esta suposición de normalidad es una hipótesis exigente que a menudo no es fácil de cumplir en la realidad clínica, donde los biomarcadores suelen tener distribuciones asimétricas. Un ejemplo ilustrativo es el cáncer de próstata, en el cual el PSA se presenta como un marcador crucial en su detección y seguimiento, mostrando una evidente asimetría [19].

Pepe et al. [83, 82] abordaron esta limitación de suposición de normalidad proponiendo un enfoque no paramétrico. Estos autores desarrollan un modelo lineal que disminuye el número de parámetros a estimar gracias a la propiedad de invarianza de la curva ROC. Además, aprovechando esta propiedad, proponen un método heurístico de búsqueda discreta de los parámetros óptimos en 201 valores equidistantes entre -1 y 1. Esto hace que la estimación sea computacionalmente abordable únicamente con un

número reducido de marcadores. Sin embargo, cuando el número de biomarcadores es mayor, la complejidad computacional se eleva al ser de tipo exponencial. Para abordar esta limitación, los autores sugieren aplicar métodos de paso a paso, donde se busca la combinación parcial óptima de variables, introduciendo una nueva variable en cada paso. Esto simplifica el problema a la estimación de un solo coeficiente en cada paso.

Otros autores, como Liu et al. [56], abordaron las limitaciones anteriores mediante el desarrollo de otro enfoque no paramétrico (enfoque min-max) que transforma el problema original considerando únicamente los valores mínimo y máximo de los biomarcadores originales, calculados para cada paciente. La idea de utilizar el máximo y el mínimo como nuevos biomarcadores puede tener su fundamento, puesto que, por definición, estos biomarcadores alcanzan la máxima sensibilidad y especificidad, respectivamente. Esto los convierte en opciones plausibles en situaciones donde se prioriza la sensibilidad (o especificidad). Sin embargo, en la práctica, es poco común maximizar una de estas métricas sin tener en cuenta la otra; suele ser más común optimizar métricas balanceadas como el AUC o el índice de Youden. Estos resultados impulsaron a los autores al desarrollo de su propuesta, que busca un compromiso entre el biomarcador máximo y mínimo. Concretamente, se centra en la estimación de la combinación lineal de estos biomarcadores que maximiza el AUC. Por tanto, es un enfoque eficiente computacionalmente, con una carga computacional similar independientemente del número de biomarcadores originales. Sin embargo, a pesar de las expectativas de los autores basadas en los resultados previos, estudios adicionales han demostrado que en algunos escenarios, el enfoque min–max tiende a alcanzar menor optimalidad en comparación con otros métodos que utilizan información de todos los biomarcadores, como los enfoques paso a paso [29, 48, 47].

Algoritmos propuestos

Nuestros trabajos (capítulos 5 y 6) parten de estos estudios previos de la literatura, con el propósito de aprovechar sus ventajas y abordar sus limitaciones. En concreto, en esta tesis, se presentan dos nuevos enfoques no paramétricos para la combinación lineal de biomarcadores continuos bajo optimización del índice de Youden: un enfoque paso a paso [4] y los enfoques denominados Min-Max-Median, Min-Max-IQR [8, 7]. Para determinar la idoneidad de nuestros enfoques propuestos, fueron comparados con otros métodos del estado del arte mediante un proceso de validación, que permitió evitar el sobreajuste y garantizar la generalización adecuada del modelo. Nuestros estudios evaluaron una amplia gama de enfoques, incluyendo técnicas estándar para problemas de clasificación como la regresión logística, enfoques como el min-max adaptado para la maximización del índice de Youden, enfoques paso a paso, estimaciones paramétricas

y no paramétricas del índice de Youden (formuladas en la sección 2.2), y modelos de Machine Learning (ML) como el XGBoost [21]. Además, se consideraron un amplio rango de escenarios simulados y reales, lo que permitió extraer conclusiones y pautas para seleccionar el algoritmo óptimo según el contexto. Nuestros estudios aportan nuevos resultados a la literatura y confirman algunas conclusiones de estudios previos que optimizaban métricas derivadas de la curva ROC. Es importante destacar que varios de estos estudios previos no utilizaron un proceso de validación, lo que podría haber influido en sus resultados debido a un posible sobreajuste y falta de generalización. Por ende, nuestro estudio, además de generar nuevos resultados, ha contribuido a reafirmar algunas conclusiones extraídas por otros autores, que optimizaban métricas derivadas de la curva ROC, y cuyos resultados debían ser interpretados con cautela y validados.

Nuestro enfoque paso a paso, presentado en el capítulo 5, sigue las sugerencias de Pepe et al. [83, 82]. El enfoque consiste en seleccionar la mejor combinación lineal de dos variables en cada paso, incluyendo una nueva variable en cada iteración. Otros autores, como Yin y Tian [124], también propusieron enfoques paso a paso bajo la maximización del índice de Youden. Sin embargo, a diferencia de nuestro enfoque, establecen un orden de entrada de las variables en cada paso y no consideran los empates. Es decir, no contemplan todas las combinaciones lineales óptimas que, en ocasiones, pueden alcanzar el mismo máximo índice de Youden. Esto hace que nuestro enfoque sea más robusto debido a que, a pesar de ser optimizaciones parciales, permite una mayor flexibilidad en la búsqueda de la mejor combinación lineal, aumentando la probabilidad de encontrar una combinación lineal más óptima.

Los resultados de nuestro estudio indican que nuestro enfoque paso a paso supera a otros en escenarios simulados con distribuciones marginales no normales y en el conjunto de datos de próstata, donde los biomarcadores, como el PSA, presentan asimetría. Esto sugiere que en escenarios no normales y con variables asimétricas, nuestro enfoque paso a paso es una opción adecuada. Estos resultados son consistentes con los reportados por otros autores, como Yin y Tian [124], quienes también sugirieron utilizar su algoritmo de paso a paso en escenarios no normales, frente a otros enfoques como el enfoque min-max o los enfoques paramétricos y no paramétricos del índice de Youden. Además, nuestro estudio demostró que, en general, nuestro algoritmo paso a paso superaba al enfoque de Yin y Tian, lo que refuerza nuestras conclusiones. Bajo maximización del AUC, Esteban et al. [29] encontraron que el algoritmo paso a paso superaba a la regresión logística en escenarios simulados con variables asimétricas.

Para los conjuntos de datos reales, el enfoque no paramétrico de tipo Kernel y nuestro algoritmo paso a paso mostraron rendimientos comparables. Sin embargo, el

enfoque no paramétrico de tipo Kernel tiene la desventaja de estar restringido a la función kernel elegida y, especialmente, al ancho de banda escogido. En escenarios con distribuciones normales, tanto la regresión logística como el enfoque paramétrico bajo normalidad multivariante, mostraron un rendimiento comparable y adecuado, en términos generales. Sin embargo, en escenarios con distribuciones no normales o variables asimétricas, el rendimiento del enfoque paramétrico se deterioró, como era de esperar. Bajo la suposición de normalidad multivariante, Su y Liu [108] ofrecen una solución óptima con AUC máximo. Para el índice de Youden, el enfoque paramétrico (2.30)-(2.33) se consideraría el equivalente. En este sentido, nuestros resultados concuerdan con los de Ma et al. [63], quienes se centraron en la maximización del pAUC y demostraron que el enfoque de Su y Liu y la regresión logística funcionaban de manera similar en escenarios normales, especialmente cuando las matrices de varianza y covarianza eran iguales, y que la regresión logística tenía un mejor rendimiento en escenarios con datos sesgados.

En nuestro estudio, el método min-max [56] demostró ser más efectivo en escenarios donde los biomarcadores tenían la misma capacidad predictiva, superando al resto de algoritmos cuando las matrices de covarianza variaban entre la población enferma y sana. Ma et al. [63] llegaron a conclusiones similares en su estudio bajo maximización del pAUC. Sin embargo, a pesar de su mejor rendimiento en estos escenarios, nuestro estudio también demostró que, en general, el enfoque min-max muestra el peor rendimiento en el resto escenarios y en los datos reales. Esto sugiere que el método min-max, al considerar solo los biomarcadores mínimo y máximo, podría no proporcionar suficiente información para discriminar adecuadamente, a pesar de su ventaja computacional.

Además de los resultados sobre el rendimiento de los métodos, la exploración de escenarios con biomarcadores altamente correlacionados también permitió corroborar los resultados del estudio de Pinsky y Zhu [85]. Los autores señalaron un incremento en el rendimiento al considerar biomarcadores altamente correlacionados negativamente. Nuestros resultados en estos escenarios respaldan esta observación, donde, en general, los algoritmos demostraron un rendimiento elevado.

En resumen, en nuestro estudio [4] presentamos un algoritmo paso a paso robusto que ha sido ampliamente comparado con otros enfoques de la literatura, demostrando ser particularmente efectivo en escenarios no normales y con variables asimétricas, comunes en la práctica real donde la normalidad no siempre se cumple. Además, hemos confirmado y validado afirmaciones de otros estudios. Sin embargo, a pesar de las fortalezas del estudio que validan la robustez del enfoque propuesto, este tiene algunas limitaciones.

La principal limitación es el tiempo computacional que exige el algoritmo paso a paso propuesto puesto que, aunque es computacionalmente abordable, requiere un tiempo significativamente mayor que el resto de enfoques comparados. Esto es debido al tratamiento de empates, que puede ser más común en tamaños de muestra pequeños, y también puede aumentar a medida que el número de biomarcadores crece. En futuras investigaciones, se buscará optimizar el algoritmo para reducir su carga computacional. Esto se logrará implementando estrategias para tratar los empates de manera más eficiente, lo que permitirá equilibrar mejor el rendimiento con la complejidad computacional. El método min-max no tiene esta restricción, puesto que es eficiente independientemente del número de biomarcadores. Sin embargo, como se ha demostrado, en algunos escenarios tiene peor capacidad de discriminación que el resto de métodos que consideran toda la información de los biomarcadores.

Los resultados anteriores impulsaron el desarrollo del segundo algoritmo propuesto [8]: enfoque Min-Max-Median (MMM) o enfoque Min-Max-IQR (MMIQR), presentado en el capítulo 6 de esta tesis. Nuestros algoritmos extienden el enfoque min-max de Liu et al. [56], incorporando una nueva estadística de resumen (mediana o rango intercuartílico). El objetivo fue considerar, además de los estadísticos extremos (mínimo y máximo), un tercer parámetro que informa sobre el rendimiento del conjunto de biomarcadores. La motivación detrás de esta extensión fue mejorar la capacidad predictiva del modelo, manteniendo al mismo tiempo las ventajas en eficiencia computacional del enfoque min-max. Aunque el algoritmo propuesto es más complejo, sigue siendo computacionalmente asequible, puesto que implica la combinación lineal de solo tres variables, independientemente del número de biomarcadores originales. Si bien la estimación de la combinación lineal de los tres estadísticos (mínimo, máximo y mediana/rango intercuartílico) puede realizarse mediante cualquier método de combinación lineal, en nuestros estudios [8, 7] optamos por implementar nuestro algoritmo paso a paso propuesto, debido a los resultados obtenidos.

En la práctica clínica, la construcción de estos enfoques tiene una interpretación clínica. Esto se debe a que existe una variedad de problemas en salud en los que la combinación del mínimo y el máximo de los biomarcadores proporciona la mejor clasificación. Por ejemplo, en el cáncer de próstata, un PSA alto y un volumen de próstata bajo pueden indicar un peor diagnóstico y se sabe que la densidad del PSA, esto es, el cociente entre el PSA y el volumen de la próstata, muestra una mejor capacidad predictiva que el PSA por sí solo. Asimismo, existen otros biomarcadores como el PCA3, SelectMdx, Proclarix y 4Kscore, que también pueden ser útiles para predecir el cáncer de próstata [80]. Los enfoques derivados del min-max dan la oportunidad de elegir, entre ellos, el que toma el valor más alto y más bajo para

cada paciente, pudiendo contribuir a mejorar la capacidad de discriminación. Además, la mediana o el rango intercuartílico proporcionan información sobre la distribución de los datos, información que puede ser útil, especialmente cuando las variables no están correlacionadas.

Nuestros enfoques MMM y MMIQR fueron comparados con el enfoque min-max y otros métodos de ML ampliamente utilizados para clasificación binaria en investigación clínica, como son la regresión logística y el algoritmo XGBoost. El objetivo fue validar más allá de técnicas tradicionales e investigar si nuestros enfoques podrían ser superiores en ciertos escenarios, incluso a algoritmos que han demostrado tener resultados prometedores en el estado del arte en los últimos años.

Los estudios de comparación mostraron que los enfoques de ML eran superiores a nuestros enfoques en los escenarios simulados de biomarcadores con diferentes capacidades predictivas, así como en escenarios con diferentes distribuciones marginales. En este último caso, el algoritmo XGBoost mostró el mejor rendimiento. En cuanto a los conjuntos de datos reales, XGBoost superó al resto de algoritmos en la predicción del riesgo de mortalidad materna, mientras que la regresión logística logró el mejor rendimiento en la predicción de la distrofia de Duchenne, un problema más sencillo. En este último caso, nuestros enfoques propuestos siguieron de cerca a la regresión logística. Los resultados podrían sugerir que el algoritmo XGBoost tiende a funcionar mejor en escenarios donde los datos muestran patrones más complejos, lo cual es coherente con su capacidad conocida para capturar relaciones no lineales.

Sin embargo, nuestros enfoques superaron a los de ML en escenarios con biomarcadores con misma capacidad predictiva y diferentes correlaciones entre grupos. Además, nuestros algoritmos funcionaron mejor que el enfoque min-max cuando los biomarcadores eran independientes y en escenarios reales. Estos resultados pueden ser interpretados como se discute a continuación. La variabilidad en las capacidades predictivas de los biomarcadores puede sugerir que uno de ellos sea más efectivo para diferenciar entre los grupos de pacientes. En estos casos, podría ser importante considerar este biomarcador en la combinación. Sin embargo, en nuestro enfoque, los biomarcadores mínimo, máximo o mediano/IQR pueden corresponder a diferentes biomarcadores para cada paciente. Esto puede ser una limitación en este caso, debido a que, por ejemplo, el valor máximo podría corresponder a un biomarcador que no es el óptimo, cuyo valor podría aportar más información que el máximo. Sin embargo, cuando los biomarcadores tienen la misma capacidad predictiva, esta elección podría tener sentido, puesto que lo que posiblemente importe es el valor máximo alcanzado, y no el biomarcador en sí. En el ámbito médico, puede ser común encontrar diferentes correlaciones entre biomarcadores para el grupo de enfermos y sanos, ya que la

biología subyacente puede ser diferente entre los grupos. Además, si los biomarcadores están altamente correlacionados, sus valores tienden a cambiar juntos, y considerar el máximo y mínimo podría ser suficiente. En cambio, en escenarios con biomarcadores no correlacionados, la información de los valores mínimo y máximo podría ser insuficiente, y agregar información sobre la mediana, por ejemplo, puede mejorar el rendimiento del modelo, puesto que proporciona una visión más completa de la distribución de los datos.

En resumen, presentamos nuevos enfoques no paramétricos (MMM, MMIQR) que fueron comparados con otros métodos del estado del arte que capturan la heterogeneidad de los biomarcadores desde una perspectiva diferente [8, 7]. Nuestro enfoque reduce la dimensionalidad del problema, considerando solo tres estadísticos de resumen, demostrando un buen rendimiento en ciertos escenarios de biomarcadores con misma capacidad predictiva. Si bien existen varias técnicas que reducen la dimensionalidad del problema, nuestro enfoque lo hace desde una perspectiva diferente, permitiendo que los tres biomarcadores considerados (mínimo, máximo y mediana o rango intercuartílico) puedan corresponder a diferentes biomarcadores originales para cada paciente.

A pesar de las fortalezas de nuestros estudios [8, 7], estos también presentan algunas limitaciones. En primer lugar, todas las conclusiones derivadas están limitadas a los escenarios y algoritmos explorados. En este sentido, aunque los algoritmos XGBoost y la regresión logística han sido ampliamente utilizados en los últimos años, demostrando ser eficientes en numerosos estudios, sería valioso contemplar como trabajo futuro una comparación con otros algoritmos de ML, especialmente en escenarios donde nuestros enfoques son óptimos. Por otro lado, aunque siguieron de cerca a la regresión logística, nuestros enfoques MMM/IQR no fueron los mejores en ninguno de los conjuntos de datos reales examinados. Como línea de investigación futura, proponemos evaluar el rendimiento de nuestro enfoque en datos reales que cumplan con las condiciones de los escenarios de simulación óptimos. Además de los escenarios que combinan múltiples biomarcadores con la misma capacidad predictiva, nuestro enfoque también podría aplicarse en escenarios donde se registran mediciones repetidas de un solo biomarcador, convirtiendo la información temporal en tres medidas resumen.

El trabajo de investigación realizado en los estudios previamente discutidos dió lugar a la creación de la librería SLModels [6] en R (Anexo A), que incorpora los algoritmos propuestos (paso a paso, MMM, MMIQR), así como el enfoque min-max, bajo maximización del índice de Youden. De esta forma, se pone a libre disposición de la comunidad científica los algoritmos desarrollados. Dado un conjunto de datos, la función *SLModels* devuelve la combinación lineal óptima (coeficientes para

cada variable) que maximiza el índice de Youden, según el algoritmo seleccionado. Adicionalmente, la función devuelve el índice de Youden alcanzado y el punto de corte óptimo asociado. Esto permite dicotomizar la información en dos grupos (sanos-enfermos), lo que facilita su utilidad clínica.

Asimismo, es interesante mencionar que nuestros enfoques han sido utilizados en estudios posteriores recientemente publicados. En concreto, Neumann et al. [75] presentaron en 2023 un estudio sobre la combinación de biomarcadores del habla para analizar la patología de la Esclerosis Lateral Amiotrófica (ELA). En él, utilizaron nuestro enfoque paso a paso [4] a través de nuestra librería SLModels [6] para calcular los pesos de las variables. Por otro lado, Condurache et al. [23] publicaron un estudio en el que implementaron un método similar al nuestro, dos años más tarde (2023). En concreto, utilizaron nuestro enfoque Min-Max-Median [8] como base, aunque aplicaron una regresión logística en lugar de nuestro algoritmo paso a paso. Siguiendo un procedimiento similar al nuestro, evaluaron su rendimiento y lo compararon con la regresión logística en varios escenarios simulados. En concreto, los datos simulados fueron generados utilizando el mismo método que empleamos en nuestro estudio, tal y como mencionan los autores. Por tanto, nuestros enfoques propuestos (algoritmo paso a paso y enfoque MMM) han servido de base para el desarrollo de nuevos estudios. Esto indica un interés en nuestras propuestas, especialmente considerando el corto lapso de tiempo desde nuestra publicación hasta su publicación posterior.

Aplicaciones en medicina

Continuando la línea de investigación desarrollada en esta tesis, en el capítulo 7 se presentan el resto de trabajos que forman parte del compendio, centrados en la estimación y aplicación de modelos en problemas de clasificación en entornos clínicos reales. Estas investigaciones resultan en herramientas esenciales para enfrentar problemas clínicos tales como el diagnóstico de enfermedades, predicciones de eventos e identificación de factores de riesgo. Los trabajos exploran diversas áreas y escenarios, incluyendo la pandemia del COVID-19, la obstetricia y el desarrollo de cáncer colorrectal. Los resultados obtenidos en estos trabajos facilitan la práctica clínica y contribuyen de manera positiva a la mejora del sistema de salud, proporcionando información relevante para la toma de decisiones médicas, la gestión de recursos y/o la planificación de políticas sanitarias. En este sentido, la elección del modelo adecuado es fundamental y depende del contexto y los datos específicos en análisis. Las investigaciones realizadas han sido llevadas a cabo en colaboración con personal experto del Instituto de Investigación Sanitaria de Aragón, Hospital General de Villalba, Hospital Universitario Miguel Servet, Hospital Clínico Universitario Lozano Blesa,

Instituto Tecnológico de Aragón y Universidad de Zaragoza.

El trabajo abordado en esta tesis [5], presentado en la sección 7.1, se enmarca dentro del contexto de la pandemia del COVID-19. El objetivo principal fue desarrollar un modelo predictivo capaz de evaluar la gravedad en pacientes hospitalizados con COVID-19 (evaluando el ingreso en UCI o la mortalidad). El propósito era dotar a los profesionales de la salud de una herramienta que facilitase la toma de decisiones rápidas, particularmente útil durante una crisis sanitaria global como esta pandemia. Para lograrlo, se exploraron técnicas de ML, evaluando la discriminación, generalización y calibración de los modelos. El algoritmo XGBoost destacó como el mejor modelo obtenido. Dada la importancia del contexto sanitario, se llevó a cabo un proceso para garantizar que la herramienta final fuera práctica y beneficiosa para los profesionales en la toma de decisiones, abordando así las limitaciones encontradas en trabajos relacionados, como la utilidad clínica, entre otros aspectos. En este sentido, el análisis del punto de corte óptimo adquiere una relevancia significativa al traducir la información y las consecuencias de la dicotomización. En el estudio, se presentó un análisis de la utilidad clínica del modelo mediante la evaluación de los puntos de corte según diferentes métricas. Esto permite al clínico seleccionar el óptimo según las necesidades del sistema y objetivos clínicos, que pueden variar con el tiempo.

Estos objetivos clínicos dependen del escenario sanitario. Por ejemplo, en escenarios de screening, se puede priorizar no perder ningún caso, mientras que en situaciones que implican intervenciones invasivas, reducir los falsos positivos puede ser más beneficioso. Además del escenario, el momento sanitario es crucial para la toma de decisiones. En el contexto de la pandemia del COVID-19, se vivió una situación compleja que obligó a ralentizar y paralizar muchos procedimientos, con consecuencias en la escasez de recursos. En estas situaciones críticas, la elección del punto de corte debe considerar no solo las implicaciones clínicas, sino también los recursos disponibles. En el artículo [3] de este compendio (sección 7.4), presentamos una metodología desarrollada con el fin de optimizar la estrategia de detección del cáncer colorrectal en el programa de cribado de Aragón. Esta iniciativa surgió como respuesta a los cambios en el ámbito de la salud generados por la pandemia, los cuales conllevan una mayor demanda y una menor oferta de colonoscopias. En este estudio, se evaluaron varios escenarios para determinar el punto de corte óptimo basado en la prueba de hemoglobina fecal, centrándose en minimizar las lesiones no diagnosticadas (falsos negativos) de acuerdo con la disponibilidad estimada de colonoscopias. Este análisis resulta crucial para la mejora del sistema de salud en Aragón en lo que respecta al cribado de cáncer colorrectal. Una vez realizado nuestro estudio, el siguiente paso debe ser la realización de validaciones externas de manera que se pueda verificar si es posible considerar este

punto de corte de manera universal.

En el ámbito de la obstetricia, es habitual emplear puntos de corte basados en tablas de percentiles de crecimiento fetal para categorizar a los fetos según su peso estimado, lo que facilita la identificación de aquellos que están por debajo o por encima de un determinado percentil para su edad gestacional. Estos fetos pueden estar vinculados a un mayor riesgo de resultados perinatales adversos, por lo que la determinación de estos puntos de corte es fundamental. Los estándares de crecimiento dependen de las características de la población así como de la estructura del modelo. En este sentido, utilizar estándares diseñados para embarazos únicos en embarazos gemelares resulta inadecuado. Por consiguiente, se requiere desarrollar modelos adaptados a las características de la población específica sobre la que se aplicarán, especialmente en el caso de embarazos gemelares, que son menos frecuentes.

En el trabajo [97] (sección 7.2), desarrollamos estándares de crecimiento fetal para gestaciones gemelares por corionicidad placentaria en una población española. El estudio analizó datos de mediciones repetidas del peso estimado gestacional para cada feto y madre, lo que implicaba una violación de la suposición de independencia entre las observaciones. Para abordar esta estructura de datos, se optó por utilizar modelos lineales mixtos, por ser una técnica adecuada en este contexto de análisis. Los modelos mixtos ofrecen una predicción más ajustada al considerar efectos aleatorios, al tiempo que proporcionan una estimación más precisa de la varianza, lo cual es fundamental para la estimación de los percentiles de peso. Se compararon los modelos desarrollados con estándares europeos y estadounidenses, mostrando un buen ajuste a la población analizada. Además, se evaluó la capacidad predictiva de los modelos para detectar casos de peso fetal bajo, utilizando métricas derivadas de la curva ROC (sensibilidad y valor predictivo positivo).

Este estudio de investigación condujo al desarrollo de la librería PTwins [2] en R (Anexo B), la cual implementa el modelo creado, proporcionando el percentil del peso fetal, dado su peso estimado y edad gestacional y corionicidad. Por tanto, la estimación de los puntos de corte (percentiles) a través de modelos lineales apropiados permitió la creación de estándares de crecimiento fetal para embarazos gemelares, resultando en una herramienta clínica válida y útil para profesionales expertos.

La práctica extendida del desarrollo de enfoques lineales también se aplica al tratamiento de factores genéticos mediante la construcción de puntuaciones de riesgo genético (GRS). Estas puntuajes combinan estos factores de forma lineal, ofreciendo así una capacidad predictiva mejorada. El análisis del valor predictivo de estos factores a través del GRS ha sido un área de gran interés y avance en la investigación médica en los últimos años. Esto se debe a que enfermedades como el cáncer colorrectal se reconocen

ampliamente como resultado de complejas interacciones entre factores ambientales y genéticos, además de los avances y reducción del costo de las pruebas genéticas. En este contexto se centra el trabajo presentado en esta tesis [34] (sección 7.3). En esta investigación, evaluamos la capacidad predictiva del GRS en el desarrollo de adenomas colorrectales, utilizando modelos de regresión logística tanto para su estimación como para analizar su asociación. Los resultados demostraron que el GRS es un factor de riesgo significativo para el desarrollo de adenomas colorrectales. Estos hallazgos pueden tener importantes implicaciones para la práctica clínica, dado que sugieren que los GRS podrían utilizarse para categorizar el riesgo de cáncer colorrectal o en programas de detección, proporcionando herramientas de diagnóstico más precisas. Por lo tanto, el uso de enfoques lineales sigue siendo común en problemas clínicos como este, debido a su facilidad de interpretación y su efectividad en la práctica clínica. Específicamente, el GRS ofrece un valor fácilmente interpretable por los clínicos, puesto que comúnmente se representa como la suma de alelos de riesgo y puede ser estratificado en categorías de riesgo genético.

Conclusiones

En resumen, los estudios presentados en esta memoria siguen la línea de investigación de la tesis centrada en la estimación de modelos para problemas de clasificación en el campo de la salud. Específicamente, el trabajo principal de la tesis ha dado lugar a la introducción en la literatura de nuevos enfoques no paramétricos para la combinación de biomarcadores continuos mediante la optimización del índice de Youden. Estos enfoques abordan limitaciones previamente observadas en la literatura. La elección de esta métrica de optimización se basa en su idoneidad y neutralidad cuando no hay consenso sobre la priorización de una clase sobre otra, convirtiéndola en una métrica clínicamente útil que no solo sirve como indicador del rendimiento del modelo, sino también como criterio para la selección del punto de corte óptimo. El análisis realizado implicó una comparación exhaustiva de métodos y escenarios, lo que permitió demostrar que nuestros enfoques propuestos pueden ser opciones adecuadas en determinadas situaciones, superiores en capacidad de discriminación a otros métodos. Esta comparación ha sido extensamente validada y, hasta donde sabemos, no hay estudios que hayan comparado un conjunto tan amplio de escenarios y métodos con diversas características en su construcción.

Además de los enfoques propuestos, el trabajo de la tesis también abarca la estimación y aplicación de modelos en diversas áreas de la salud, abordando problemas clínicos reales como la predicción de la gravedad del COVID-19, la optimización de estrategias de cribado del cáncer colorrectal, el desarrollo de estándares de crecimiento

fetal para embarazos gemelares y la evaluación de la asociación entre el riesgo genético y el desarrollo de adenomas colorrectales. Estos trabajos han generado herramientas y metodologías con impacto directo en la práctica clínica y la calidad de la atención sanitaria, validadas en colaboración con instituciones y hospitales.

Aunque el foco principal de la investigación de esta tesis fue construir modelos con una mayor capacidad predictiva para ofrecer predicciones más precisas, otro aspecto destacado fue la utilidad clínica. En este sentido, la aplicación de enfoques interpretables, desarrollo de herramientas y el análisis del punto de corte óptimo también han sido aspectos importantes en los trabajos científicos que conforman este compendio, abordándose de diversas maneras según los diferentes objetivos. Además, el trabajo realizado se ha traducido no solo en herramientas clínicas, sino también en librerías de R que incluyen los enfoques propuestos y los modelos desarrollados, facilitando su aplicación. Estas librerías se han puesto a libre disposición de la comunidad científica y los profesionales interesados en repositorios de acceso público (CRAN).

En conclusión, los trabajos presentados en este compendio representan una contribución significativa al campo de la bioestadística, proporcionando herramientas y conocimientos que podrían resultar en mejoras tangibles en la atención médica y la salud pública, dentro del ámbito de la salud. Asimismo, se espera seguir investigando en esta dirección, abordando las limitaciones identificadas y mejorando y ampliando los algoritmos propuestos para abarcar más escenarios y posibilidades.

Bibliografía

[1] Agresti, A. and Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126.

[2] Aznar, R., Esteban, L. M., Sanz, G., and Saviron, R. (2019). *PTwins: Percentile Estimation of Fetal Weight for Twins by Chorionicity*. R package version 0.1.1.

[3] Aznar-Gimeno, R., Carrera-Lasfuentes, P., del Hoyo-Alonso, R., Doblaré, M., and Lanas, Á. (2021a). Evidence-based selection on the appropriate fit cut-off point in crc screening programs in the covid pandemic. *Frontiers in medicine*, 8:712040.

[4] Aznar-Gimeno, R., Esteban, L. M., del Hoyo-Alonso, R., Borque-Fernando, Á., and Sanz, G. (2022a). A stepwise algorithm for linearly combining biomarkers under youden index maximization. *Mathematics*, 10(8):1221.

[5] Aznar-Gimeno, R., Esteban, L. M., Labata-Lezaun, G., del Hoyo-Alonso, R., Abadia-Gallego, D., Paño-Pardo, J. R., Esquillor-Rodrigo, M. J., Lanas, Á., and Serrano, M. T. (2021b). A clinical decision web to predict icu admission or death for patients hospitalised with covid-19 using machine learning algorithms. *International Journal of Environmental Research and Public Health*, 18(16):8677.

[6] Aznar-Gimeno, R., Esteban, L. M., Sanz, G., and del Hoyo-Alonso, R. (2022b). *SLModels: Stepwise Linear Models for Binary Classification Problems under Youden Index Optimisation*. R package version 0.1.2.

[7] Aznar-Gimeno, R., Esteban, L. M., Sanz, G., and del Hoyo-Alonso, R. (2023). Comparing the min–max–median/iqr approach with the min–max approach, logistic regression and xgboost, maximising the youden index. *Symmetry*, 15(3):756.

[8] Aznar-Gimeno, R., Esteban, L. M., Sanz, G., del Hoyo-Alonso, R., and Savirón-Cornudella, R. (2021c). Incorporating a new summary statistic into the min–max approach: a min–max–median, min–max–iqr combination of biomarkers for maximising the youden index. *Mathematics*, 9(19):2497.

[9] Aznar-Gimeno, R., Paño-Pardo, J. R., Esteban, L. M., Labata-Lezaun, G., Esquillor-Rodrigo, M. J., Lanas, A., Abadía-Gallego, D., Diez-Fuertes, F., Tellería-Orriols, C., del Hoyo-Alonso, R., et al. (2021d). Changes in severity, mortality, and virus genome among a spanish cohort of patients hospitalized with sars-cov-2. *Scientific Reports*, 11(1):18844.

[10] Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of mathematical psychology*, 12(4):387–415.

[11] Bentéjac, C., Csörgő, A., and Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54:1937–1967.

[12] Berkson, J. (1947). "cost-utility" as a measure of the efficiency of a test. *Journal of the American Statistical Association*, 42(238):246–255.

[13] Bertail, P., Clémençcon, S., and Vayatis, N. (2008). On bootstrapping the roc curve. *Advances in Neural Information Processing Systems*, 21.

[14] Böhning, D., Holling, H., and Patilea, V. (2011). A limitation of the diagnostic-odds ratio in determining an optimal cut-off value for a continuous diagnostic test. *Statistical methods in medical research*, 20(5):541–550.

[15] Borque, Á., Rubio-Briones, J., Esteban, L. M., Sanz, G., Domínguez-Escrig, J., Ramírez-Backhaus, M., Calatrava, A., and Solsona, E. (2014). Implementing the use of nomograms by choosing threshold points in predictive models: 2012 updated p artin t ables vs a e uropean predictive nomogram for organ-confined disease in prostate cancer. *BJU international*, 113(6):878–886.

[16] Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 26(2):211–243.

[17] Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical science*, 16(2):101–133.

[18] Cai, T. and Moskowitz, C. S. (2004). Semi-parametric estimation of the binormal roc curve for a continuous diagnostic test. *Biostatistics*, 5(4):573–586.

[19] Capitanio, U., Perrotte, P., Zini, L., Suardi, N., Antebi, E., Cloutier, V., Jeldres, C., Shariat, S. F., Duclos, A., Arjane, P., et al. (2009). Population-based analysis of normal total psa and percentage of free/total psa values: results from screening cohort. *Urology*, 73(6):1323–1327.

[20] Carayol, J., Tores, F., König, I. R., Hager, J., and Ziegler, A. (2010). Evaluating diagnostic accuracy of genetic profiles in affected offspring families. *Statistics in medicine*, 29(22):2359–2368.

[21] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

[22] Clopper, C. J. and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413.

[23] Condurache, I.-A. and Bolboacă, S. D. (2023). Min-max, min-max-median, and min-max-iqr in deciding optimal diagnostic thresholds: Performances of a logistic regression approach on simulated and real data. *Applied Medical Informatics*, 45(3).

[24] de Hond, A. A., Kant, I. M., Honkoop, P. J., Smith, A. D., Steyerberg, E. W., and Sont, J. K. (2022). Machine learning did not beat logistic regression in time series prediction for severe asthma exacerbations. *Scientific Reports*, 12(1):20363.

[25] DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44:837–845.

[26] Demidenko, E. (2012). Confidence intervals and bands for the binormal roc curve revisited. *Journal of applied statistics*, 39(1):67–79.

[27] Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer.

[28] Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman and Hall/CRC.

[29] Esteban, L. M., Sanz, G., and Borque, A. (2011). A step-by-step algorithm for combining diagnostic tests. *Journal of Applied Statistics*, 38(5):899–911.

[30] Faraggi, D. and Reiser, B. (2002). Estimation of the area under the roc curve. *Statistics in medicine*, 21(20):3093–3106.

[31] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

[32] Fluss, R., Faraggi, D., and Reiser, B. (2005). Estimation of the youden index and its associated cutoff point. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 47(4):458–472.

[33] Franco, M. and Vivo, J. M. (2007). Análisis de curvas roc. principios básicos y aplicaciones. *Madrid: La Muralla*.

[34] Gargallo-Puyuelo, C. J., Aznar-Gimeno, R., Carrera-Lasfuentes, P., Lanas, A., Ferrandez, A., Quintero, E., Carrillo, M., Alonso-Abreu, I., Esteban, L. M., de la Vega Rodrigalvarez-Chamarro, M., et al. (2022). Predictive value of genetic risk scores in the development of colorectal adenomas. *Digestive Diseases and Sciences*, 67(8):4049–4058.

[35] Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J., and Bossuyt, P. M. (2003). The diagnostic odds ratio: a single indicator of test performance. *Journal of clinical epidemiology*, 56(11):1129–1135.

[36] Gonçalves, L., Subtil, A., Oliveira, M. R., and de Zea Bermudez, P. (2014). Roc curve estimation: An overview. *REVSTAT-Statistical journal*, 12(1):1–20.

[37] Green, D. M., Swets, J. A., et al. (1966). *Signal detection theory and psychophysics*, volume 1. Wiley New York.

[38] Hall, P., Hyndman, R. J., and Fan, Y. (2004). Nonparametric confidence intervals for receiver operating characteristic curves. *Biometrika*, 91(3):743–750.

[39] Hall, P. G. and Hyndman, R. J. (2003). Improved methods for bandwidth selection when estimating roc curves. *Statistics & Probability Letters*, 64(2):181–189.

[40] Hanley, J. A. (1988). The robustness of the "binormal" assumptions used in fitting roc curves. *Medical decision making*, 8(3):197–203.

[41] Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

[42] Hanley, J. A. and McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3):839–843.

[43] Hsieh, F. and Turnbull, B. W. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The annals of statistics*, 24(1):25–40.

[44] Hsu, M.-J. and Hsueh, H.-M. (2013). The linear combinations of biomarkers which maximize the partial area under the roc curves. *Computational Statistics*, 28:647–666.

[45] Jokiel-Rokita, A. and Pulit, M. (2013). Nonparametric estimation of the roc curve based on smoothed empirical distribution functions. *Statistics and Computing*, 23:703–712.

[46] Jung, K., Lee, J., Gupta, V., and Cho, G. (2019). Comparison of bootstrap confidence interval methods for gsca using a monte carlo simulation. *Frontiers in psychology*, 10:472295.

[47] Kang, L., Liu, A., and Tian, L. (2016). Linear combination methods to improve diagnostic/prognostic accuracy on future observations. *Statistical methods in medical research*, 25(4):1359–1380.

[48] Kang, L., Xiong, C., Crane, P., and Tian, L. (2013). Linear combinations of biomarkers to improve diagnostic accuracy with three ordinal diagnostic categories. *Statistics in medicine*, 32(4):631–643.

[49] Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

[50] Larsson, A., Berg, J., Gellerfors, M., and Gerdin Wärnberg, M. (2021). The advanced machine learner xgboost did not reduce prehospital trauma mistriage compared with logistic regression: a simulation study. *BMC medical informatics and decision making*, 21(1):1–9.

[51] Li, C., Chen, J., and Qin, G. (2019). Partial youden index and its inferences. *Journal of Biopharmaceutical Statistics*, 29(2):385–399.

[52] Li, D.-l., Shen, F., Yin, Y., Peng, J.-x., and Chen, P.-y. (2013). Weighted youden index and its two-independent-sample comparison based on weighted sensitivity and specificity. *Chinese medical journal*, 126(6):1150–1154.

[53] Littell, R. C. and Milliken, G. A. (2006). *SAS for mixed models*.

[54] Liu, A. and Schisterman, E. F. (2003). Comparison of diagnostic accuracy of biomarkers with pooled assessments. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 45(5):631–644.

[55] Liu, A., Schisterman, E. F., and Zhu, Y. (2005). On linear combinations of biomarkers to improve diagnostic accuracy. *Statistics in medicine*, 24(1):37–47.

[56] Liu, C., Liu, A., and Halabi, S. (2011). A min–max combination of biomarkers to improve diagnostic accuracy. *Statistics in medicine*, 30(16):2005–2014.

[57] Liu, X. (2012). Classification accuracy and cut point selection. *Statistics in medicine*, 31(23):2676–2686.

[58] López Ratón, M. (2015). *Optimal cutoff points for classification in diagnostic studies: new contributions and software development*. PhD thesis.

[59] López-Ratón, M., Cadarso-Suárez, C., Molanes-López, E. M., and Letón, E. (2016). Confidence intervals for the symmetry point: An optimal cutpoint in continuous diagnostic tests. *Pharmaceutical Statistics*, 15(2):178–192.

[60] López Ratón, M., Molanes López, E. M., Letón, E., and Cadarso Suárez, C. M. (2017). Gsympoint: An r package to estimate the generalized symmetry point, an optimal cut-off point for binary classification in continuous diagnostic tests.

[61] López-Ratón, M., Rodríguez-Álvarez, M. X., Cadarso-Suárez, C., and Gude-Sampedro, F. (2014). Optimalcutpoints: an r package for selecting optimal cutpoints in diagnostic tests. *Journal of statistical software*, 61:1–36.

[62] Lusted, L. B. (1971). Signal detectability and medical decision-making: Signal detectability studies help radiologists evaluate equipment systems and performance of assistants. *Science*, 171(3977):1217–1219.

[63] Ma, H., Halabi, S., Liu, A., et al. (2019). On the use of min-max combination of biomarkers to maximize the partial area under the roc curve. *Journal of probability and statistics*, 2019.

[64] Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 18:50–60.

[65] Mantalos, P. and Zografos, K. (2008). Interval estimation for a binomial proportion: a bootstrap approach. *Journal of Statistical Computation and Simulation*, 78(12):1251–1265.

[66] Martínez-Camblor, P., Pérez-Fernández, S., and Corral, N. (2018). Efficient nonparametric confidence bands for receiver operating-characteristic curves. *Statistical Methods in Medical Research*, 27(6):1892–1908.

[67] Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

[68] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

[69] Metz, C. E. (1978). Basic principles of roc analysis. In *Seminars in nuclear medicine*, volume 8, pages 283–298. Elsevier.

[70] Metz, C. E., Herman, B. A., and Shen, J.-H. (1998). Maximum likelihood estimation of receiver operating characteristic (roc) curves from continuously-distributed data. *Statistics in medicine*, 17(9):1033–1053.

[71] Metz, C. E. and Kronman, H. B. (1980). Statistical significance tests for binormal roc curves. *Journal of Mathematical Psychology*, 22(3):218–243.

[72] Metz, C. E., Wang, P.-L., and Kronman, H. B. (1984). A new approach for testing the significance of differences between roc curves measured from correlated data. In *Information Processing in Medical Imaging: Proceedings of the 8th conference, Brussels, 29 August–2 September 1983*, pages 432–445. Springer.

[73] Nahm, F. S. (2022). Receiver operating characteristic curve: overview and practical use for clinicians. *Korean journal of anesthesiology*, 75(1):25.

[74] Navarro, C. L. A., Damen, J. A., van Smeden, M., Takada, T., Nijman, S. W., Dhiman, P., Ma, J., Collins, G. S., Bajpai, R., Riley, R. D., et al. (2023). Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. *Journal of clinical epidemiology*, 154:8–22.

[75] Neumann, M., Kothare, H., and Ramanarayanan, V. (2023). Combining Multiple Multimodal Speech Features into an Interpretable Index Score for Capturing Disease Progression in Amyotrophic Lateral Sclerosis. In *Proc. INTERSPEECH 2023*, pages 2353–2357.

[76] Obuchowski, N. A. and McCLISH, D. K. (1997). Sample size determination for diagnostic accuracy studies involving binormal roc curve indices. *Statistics in medicine*, 16(13):1529–1542.

[77] Oehlert, G. W. (1992). A note on the delta method. *The American Statistician*, 46(1):27–29.

[78] Pappada, S. M. (2021). Machine learning in medicine: It has arrived, let's embrace it. *Journal of Cardiac Surgery*, 36(11):4121–4124.

[79] Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076.

[80] Pastor-Navarro, B., Rubio-Briones, J., Borque-Fernando, Á., Esteban, L. M., Dominguez-Escrig, J. L., and López-Guerrero, J. A. (2021). Active surveillance in prostate cancer: role of available biomarkers in daily practice. *International Journal of Molecular Sciences*, 22(12):6266.

[81] Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford university press.

[82] Pepe, M. S., Cai, T., and Longton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics*, 62(1):221–229.

[83] Pepe, M. S. and Thompson, M. L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics*, 1(2):123–140.

[84] Perkins Neil, J. and Schisterman Enrique, F. (2006). The inconsistency of "optimal" cut-points using two roc based criteria. *Am. J. Epidemiol*, 163:670–675.

[85] Pinsky, P. F. and Zhu, C. S. (2011). Building multi-marker algorithms for disease prediction—the role of correlations among markers. *Biomarker insights*, 6:BMI–S7513.

[86] Platt, R. W., Hanley, J. A., and Yang, H. (2000). Bootstrap confidence intervals for the sensitivity of a quantitative diagnostic test. *Statistics in medicine*, 19(3):313–322.

[87] Qin, G. and Hotilovac, L. (2008). Comparison of non-parametric confidence intervals for the area under the roc curve of a continuous-scale diagnostic test. *Statistical methods in medical research*, 17(2):207–221.

[88] Qin, G. and Zhou, X.-H. (2006). Empirical likelihood inference for the area under the roc curve. *Biometrics*, 62(2):613–622.

[89] Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, 12:1–8.

[90] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The annals of mathematical statistics*, 27:832–837.

[91] Rota, M. and Antolini, L. (2014). Finding the optimal cut-point for gaussian and gamma distributed biomarkers. *Computational Statistics & Data Analysis*, 69:1–14.

[92] Rubio-Briones, J., Borque, A., Esteban, L. M., Casanova, J., Fernandez-Serra, A., Rubio, L., Casanova-Salas, I., Sanz, G., Domínguez-Escrig, J., Collado, A., et al. (2015). Optimizing the clinical utility of pca3 to diagnose prostate cancer in initial prostate biopsy. *BMC cancer*, 15:1–8.

[93] Rücker, G. and Schumacher, M. (2010). Summary roc curve based on a weighted youden index for selecting an optimal cutpoint in meta-analysis of diagnostic accuracy. *Statistics in medicine*, 29(30):3069–3078.

[94] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, 2(3):160.

[95] Savirón-Cornudella, R., Esteban, L. M., Aznar-Gimeno, R., Dieste-Pérez, P., Pérez-López, F. R., Campillos, J. M., Castán-Larraz, B., Sanz, G., and Tajada-Duaso, M. (2021a). Prediction of late-onset small for gestational age and fetal growth restriction by fetal biometry at 35 weeks and impact of ultrasound–delivery interval: Comparison of six fetal growth standards. *Journal of Clinical Medicine*, 10(13):2984.

[96] Savirón-Cornudella, R., Esteban, L. M., Aznar-Gimeno, R., Dieste Pérez, P., Pérez-López, F. R., Castán-Larraz, B., Sanz, G., and Tajada-Duaso, M. (2021b). Prediction of large for gestational age by ultrasound at 35 weeks and impact of ultrasound-delivery interval: Comparison of 6 standards. *Fetal Diagnosis and Therapy*, 48(1):15–23.

[97] Savirón-Cornudella, R., Esteban, L. M., Aznar-Gimeno, R., Pérez-López, F. R., Ezquerro, M. C., Pérez, P. D., Maza, J. M. C., Sanz, G., Larraz, B. C., and Tajada-Duaso, M. (2020). A cohort study of fetal growth in twin pregnancies by chorionicity: comparison with european and american standards. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 253:238–248.

[98] Schisterman, E. F. and Perkins, N. (2007). Confidence intervals for the youden index and corresponding optimal cut-point. *Communications in Statistics—Simulation and Computation®*, 36(3):549–563.

[99] Sen, P. K. (1960). On some convergence properties of ustatistics. *Calcutta Statistical Association Bulletin*, 10(1-2):1–18.

[100] Shan, G. (2015). Improved confidence intervals for the youden index. *PloS one*, 10(7):e0127272.

[101] Sheather, S. J. (2004). Density estimation. *Statistical science*, pages 588–597.

[102] Shehab, M., Abualigah, L., Shambour, Q., Abu-Hashem, M. A., Shambour, M. K. Y., Alsalibi, A. I., and Gandomi, A. H. (2022). Machine learning in medical applications: A review of state-of-the-art methods. *Computers in Biology and Medicine*, 145:105458.

[103] Sidey-Gibbons, J. A. and Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical introduction. *BMC medical research methodology*, 19:1–18.

[104] Silverman, B. W. (1986). *Density estimation for statistics and data analysis (Vol. 26)*. CRC Press.

[105] Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):3940–3941.

[106] Somoza, E. and Mossman, D. (1991). "biological markers" and psychiatric diagnosis: risk-benefit balancing using roc analysis. *Biological psychiatry*, 29(8):811–826.

[107] Stefanescu, C., Berger, V. W., and Hershberger, S. (2005). Yates's continuity correction. *Encyclopaedia of Statistics in Behavioural Science*, 4:2127–2129.

[108] Su, J. Q. and Liu, J. S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association*, 88(424):1350–1355.

[109] Sweets, J. (1996). Signal detection theory and roc analysis in psychology and diagnostics.

[110] Swets, J. A. (1986). Form of empirical rocs in discrimination and diagnostic tasks: implications for theory and measurement of performance. *Psychological bulletin*, 99(2):181.

[111] Unal, I. (2017). Defining an optimal cut-point value in roc analysis: an alternative approach. *Computational and mathematical methods in medicine*, 2017.

[112] Venkatraman, E. (2000). A permutation test to compare receiver operating characteristic curves. *Biometrics*, 56(4):1134–1138.

[113] Venkatraman, E. S. and Begg, C. B. (1996). A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika*, 83(4):835–848.

[114] Verbeke, G., Molenberghs, G., and Verbeke, G. (1997). *Linear mixed models for longitudinal data*. Springer.

[115] Vilar, J. M. (1991). Estimación no paramétrica de la función de distribución. *Qüestiió*, 15(1):3–20.

[116] Volovici, V., Syn, N. L., Ercole, A., Zhao, J. J., and Liu, N. (2022). Steps to avoid overuse and misuse of machine learning in clinical research. *Nature Medicine*, 28(10):1996–1999.

[117] Walker, S. H. and Duncan, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54(1-2):167–179.

[118] Wallis, S. (2013). Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. *Journal of quantitative linguistics*, 20(3):178–208.

[119] Walsh, S. J. (1997). Limitations to the robustness of binormal roc curves: effects of model misspecification and location of decision thresholds on bias, precision, size and power. *Statistics in medicine*, 16(6):669–679.

[120] Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

[121] Yan, Q., Bantis, L. E., Stanford, J. L., and Feng, Z. (2018). Combining multiple biomarkers linearly to maximize the partial area under the roc curve. *Statistics in medicine*, 37(4):627–642.

[122] Yin, J. and Tian, L. (2014a). Joint confidence region estimation for area under roc curve and youden index. *Statistics in medicine*, 33(6):985–1000.

[123] Yin, J. and Tian, L. (2014b). Joint inference about sensitivity and specificity at the optimal cut-off point associated with youden index. *Computational Statistics & Data Analysis*, 77:1–13.

[124] Yin, J. and Tian, L. (2014c). Optimal linear combinations of multiple diagnostic biomarkers based on youden index. *Statistics in medicine*, 33(8):1426–1440.

[125] Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1):32–35.

[126] Yu, W. and Park, T. (2015). Two simple algorithms on linear combination of multiple biomarkers to maximize partial area under the roc curve. *Computational Statistics & Data Analysis*, 88:15–27.

[127] Zhou, H. and Qin, G. (2012). New nonparametric confidence intervals for the youden index. *Journal of biopharmaceutical statistics*, 22(6):1244–1257.

[128] Zhou, X.-H. and Harezlak, J. (2002). Comparison of bandwidth selection methods for kernel smoothing of roc curves. *Statistics in medicine*, 21(14):2045–2055.

[129] Zhou, X.-H., Obuchowski, N. A., and McClish, D. K. (2011). *Statistical methods in diagnostic medicine*, volume 712. John Wiley & Sons.

[130] Zhou, X.-H. and Qin, G. (2005). Improved confidence intervals for the sensitivity at a fixed level of specificity of a continuous-scale diagnostic test. *Statistics in medicine*, 24(3):465–477.

[131] Zou, K. H., Hall, W. J., and Shapiro, D. E. (1997). Smooth non-parametric receiver operating characteristic (roc) curves for continuous diagnostic tests. *Statistics in medicine*, 16(19):2143–2156.

[132] Zou, K. H., Tempany, C. M., Fielding, J. R., and Silverman, S. G. (1998). Original smooth receiver operating characteristic curve estimation from continuous data: statistical methods for analyzing the predictive value of spiral ct of ureteral stones. *Academic radiology*, 5(10):680–687.

[133] Zweig, M. H. and Campbell, G. (1993). Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry*, 39(4):561–577.

Anexos

Anexos A

Librería R. SLModels

Package 'SLModels'

October 12, 2022

Type Package

Title Stepwise Linear Models for Binary Classification Problems under Youden Index Optimisation

Version 0.1.2

Depends stats, ROCR

Maintainer Rocio Aznar-Gimeno <raznar@itainnova.es>

Description Stepwise models for the optimal linear combination of continuous variables in binary classification problems under Youden Index optimisation. Information on the models implemented can be found at Aznar-Gimeno et al. (2021) <doi:10.3390/math9192497>.

License GPL-3

Encoding UTF-8

NeedsCompilation no

Author Rocio Aznar-Gimeno [aut, cre] (<https://orcid.org/0000-0003-1415-146X>),
Luis Mariano Esteban [aut] (<https://orcid.org/0000-0002-3007-302X>),
Gerardo Sanz [aut] (<https://orcid.org/0000-0002-6474-2252>),
Rafael del Hoyo-Alonso [aut] (<https://orcid.org/0000-0003-2755-5500>)

Repository CRAN

Date/Publication 2022-02-03 14:30:10 UTC

R topics documented:

`SLModels` *Stepwise Linear Models for Binary Classification Problems under Youden Index Optimisation*

Description

Stepwise models for the optimal linear combination of continuous variables in binary classification problems under Youden Index optimisation. Information on the models implemented can be found at Aznar-Gimeno et al. (2021) <doi:10.3390/math9192497>.

Usage

```
SLModels(data, algorithm="stepwise", scaling=FALSE)
```

Arguments

`data`	Data frame containing the input variables and the binary output variable. The last column must be reserved for the output variable.
`algorithm`	string; Stepwise linear model to be applied. The options are: "stepwise", "minmax", "minmaxmedian", "minmaxiqr"; default value: "stepwise".
`scaling`	boolean; if TRUE, the Min-Max Scaling is applied; if FALSE, no normalisation is applied to the input variables; default value: FALSE.

Details

The "stepwise" algorithm refers to our proposed stepwise algorithm on the original variables which is the adaptation for the maximisation of the Youden index of the one proposed by Esteban et al. (2011) <doi:10.1080/02664761003692373>. The general idea of this approach, as suggested by Pepe and Thompson (2000) <doi:10.1093/biostatistics/1.2.123>, is to follow a step by step algorithm that includes a new variable in each step, selecting the best combination (or combinations) of two variables, in terms of maximising the Youden index.

The "minmax" algorithm refers to the distribution-free min–max approach proposed by Liu et al. (2011) <doi:10.1002/sim.4238>. The idea is to reduce the order of the linear combination beforehand by considering only two markers (maximum and minimum values of all the variables/biomarkers). This algorithm was adapted in order to maximise the Youden index.

The "minmaxmedian" algorithm refers to our proposed algorithm that considers the linear combination of the following three variables: the minimum, maximum and median values of the original variables.

The "minmaxiqr" algorithm refers to our proposed algorithm that considers the linear combination of the following three variables: the minimum, maximum and interquartile range (IQR) values of the original variables.

More information on the implemented algorithms can be found in Aznar-Gimeno et al. (2021) <doi:10.3390/math9192497>.

Value

Optimal linear combination that maximises the Youden index. Specifically, the function returns the coefficients for each variable, optimal cut-off point and Youden Index achieved.

Note

The "stepwise" algorithm becomes a computationally intensive problem when the number of variables exceeds 4.

Author(s)

Rocío Aznar-Gimeno, Luis Mariano Esteban, Gerardo Sanz, Rafael del Hoyo-Alonso

References

Aznar-Gimeno, R., Esteban, L. M., Sanz, G., del-Hoyo-Alonso, R., & Savirón-Cornudella, R. (2021). Incorporating a New Summary Statistic into the Min–Max Approach: A Min–Max–Median, Min–Max–IQR Combination of Biomarkers for Maximising the Youden Index. *Mathematics*, 9(19), 2497, doi:10.3390/math9192497.

Esteban, L. M., Sanz, G., & Borque, A. (2011). A step-by-step algorithm for combining diagnostic tests. *Journal of Applied Statistics*, 38(5), 899-911, doi:10.1080/02664761003692373.

Pepe, M. S., & Thompson, M. L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics*, 1(2), 123-140, doi:10.1093/biostatistics/1.2.123.

Liu, C., Liu, A., & Halabi, S. (2011). A min–max combination of biomarkers to improve diagnostic accuracy. *Statistics in medicine*, 30(16), 2005-2014, doi:10.1002/sim.4238.

Examples

```
#Create dataframe
x1<-rnorm(100,sd =1)
x2<-rnorm(100,sd =2)
x3<-rnorm(100,sd =3)
x4<-rnorm(100,sd =4)
z <- rep(c(1,0), c(50,50))
DT<-data.frame(cbind(x1,x2,x3,x4))
data<-cbind(DT,z)

#Example 1#
SLModels(data) #default values: algorithm="stepwise", scaling=FALSE
#Example 2#
SLModels(data, algorithm="minmax") #scaling=FALSE, default value
#Example 3#
SLModels(data, algorithm="minmax", scaling=TRUE)
#Example 4#
SLModels(data, algorithm="minmaxmedian", scaling=TRUE)
#Example 5#
SLModels(data, algorithm="minmaxiqr", scaling=TRUE)
```

Index

Anexos B

Librería R. PTwins

Package 'PTwins'

October 12, 2022

Type Package

Title Percentile Estimation of Fetal Weight for Twins by Chorionicity

Version 0.1.1

Author Rocio Aznar [aut],
Luis Mariano Esteban [aut, cre],
Gerardo Sanz [aut],
Ricardo Saviron [aut]

Maintainer Luis Mariano Esteban <lmeste@unizar.es>

Description Package to Percentile estimation of fetal weight for twins by chorionicity (dichorionic-diamniotic or monochorionic-diamniotic).

License GPL-3

Encoding UTF-8

LazyData true

NeedsCompilation no

Repository CRAN

Date/Publication 2019-11-19 13:00:02 UTC

R topics documented:

`PTwins` *Percentile Estimation of Fetal Weight for Twins by Chorionicity*

Description

The PTwins function estimates the fetus weight percentile using a multilevel linear model developed from a Spanish twin cohort.

Usage

```
PTwins(weight,week,day=3,dichorionic=TRUE)
```

Arguments

weight	The fetus weight estimated by ultrasound at a gestational age (weeks) using Hadlock formula: log10(FW)=1.3596-0.00386(AC)x(FL)+0.0064(HC)+0.0006(BPD)x(AC)+ 0.0424(AC)+0.174(FL) FW: fetal weight, AC: abdominal circumference, FL: femur lenght, HC: head circumference, BPD: biparital diameter
week	The gestational age in weeks for which the Fetus weight where estimated
day	The exact day in the gestational week at which the fetus weight were estimated. For the percentile calculation, if we choose for example: week=23, day=2, this is equivalent to 163 days of GA
dichorionic	This parameter indicates if the fetus is dichorionic-diamniotic (dichorionic=TRUE) or monochorionic-monoamniotic (dichorionic=FALSE)

Details

The inputs weight, week, day or dichorionic can be a number or a vector to include several cases

Value

The returned fit object of PTwins contains the following components.

Percentile	Percentile of the fetus weight
weight	fetus weight
GA	Gestational age in exact weeks

Author(s)

Rocio Aznar, Luis Mariano Esteban, Gerardo Sanz, Ricardo Saviron

Examples

```
#Percentile estimation of a dichorionic-diamniotic fetus of 2300 grams
#of weight estimated at the 22nd week (+2 days) of gestational age.

PTwins(weight=2300,week=22,day=2,dichorionic=TRUE)

#Percentile estimation of a monochorionic-diamniotic fetus of 2300 grams
#of weight estimated at the 22nd week (+2 days) of gestational age.

PTwins(weight=2300,week=22,day=2,dichorionic=FALSE)

#Percentile estimation of a dataframe that includes 10 cases

WEIGHT<-round(rnorm(10,2100,125),digits=0)
```

```
WEEK<-sample(seq(18,36),10)
DAY<-sample(seq(0,7),10,replace=TRUE)
dichorionic<-sample(c("TRUE","FALSE"),10,replace=TRUE)
DT<-data.frame(WEIGHT,WEEK,DAY,dichorionic)

PTwins(weight=DT$WEIGHT,week=DT$WEEK,day=DT$DAY,dichorionic=DT$dichorionic)
```

Index

Apéndice

Calidad de las publicaciones y contribuciones

En este anexo se muestra el factor de impacto y la categoría o área temática de las publicaciones que conforman el compendio, así como la contribución en cada una de ellas. Los parámetros han sido obtenidos a través de la plataforma Web of Sciences.

- Aznar-Gimeno, R., Esteban, L. M., del-Hoyo-Alonso, R., Borque-Fernando, Á., & Sanz, G. (2022). A Stepwise Algorithm for Linearly Combining Biomarkers under Youden Index Maximization. Mathematics, 10(8), 1221.
    - Factor de impacto JCR 2022: 2,4. 23/329 (Q1)
    - Categoría: Mathematics
    - Contribución Rocío Aznar Gimeno: Conceptualización, metodología, software, validación, análisis formal, investigación, recursos, curación de datos, visualización, supervisión, preparación del borrador original del manuscrito, revisión, edición y escritura del manuscrito final.
- Aznar-Gimeno, R., Esteban, L. M., Sanz, G., del-Hoyo-Alonso, R., & Savirón-Cornudella, R. (2021). Incorporating a new summary statistic into the min–max approach: a min–max–median, min–max–IQR combination of biomarkers for maximising the youden index. Mathematics, 9(19), 2497.
    - Factor de impacto JCR 2021: 2,542. 21/333 (Q1)
    - Categoría: Mathematics
    - Contribución Rocío Aznar Gimeno: Conceptualización, metodología, software, validación, análisis formal, investigación, recursos, curación de datos, visualización, supervisión, preparación del borrador original del manuscrito, revisión, edición y escritura del manuscrito final.
- Aznar-Gimeno, R., Esteban, L. M., Sanz, G., & del-Hoyo-Alonso, R. (2023). Comparing the Min–Max–Median/IQR Approach with the Min–Max Approach,

Logistic Regression and XGBoost, Maximising the Youden Index. Symmetry, 15(3), 756.

- Factor de impacto JCR 2022: 2,7. 36/73 (Q2)
- Categoría: Multidisciplinary Sciences
- Contribución Rocío Aznar Gimeno: Conceptualización, metodología, software, validación, análisis formal, investigación, recursos, curación de datos, visualización, supervisión, preparación del borrador original del manuscrito, revisión, edición y escritura del manuscrito final.

- Aznar-Gimeno, R., Esteban, L. M., Labata-Lezaun, G., del-Hoyo-Alonso, R., Abadia-Gallego, D., Paño-Pardo, J. R., Esquillor-Rodrigo M. J., Lanas, Á., & Serrano, M. T. (2021). A clinical decision web to predict ICU admission or death for patients hospitalised with COVID-19 using machine learning algorithms. International Journal of Environmental Research and Public Health, 18(16), 8677.

  - Factor de impacto JCR 2021: 4,614. 100/279 - 71/210 (Q2)
  - Categoría: Environmental Sciences - Public, Environmental & Occupational Health
  - Contribución Rocío Aznar Gimeno: Metodología, software, validación, análisis formal, investigación, recursos, curación de datos, visualización, preparación del borrador original del manuscrito, revisión, edición y escritura del manuscrito final.

- Savirón-Cornudella, R., Esteban, L. M., Aznar-Gimeno, R., Pérez-López, F. R., Ezquerro, M. C., Pérez, P. D., Maza, J. M., Sanz, G., Larraz, B. C., & Tajada-Duaso, M. (2020). A cohort study of fetal growth in twin pregnancies by chorionicity: comparison with European and American standards. European Journal of Obstetrics & Gynecology and Reproductive Biology, 253, 238-248.

  - Factor de impacto JCR 2020: 2,435. 49/83 - 22/30 (Q3)
  - Categoría: Obstetrics & Gynecology - Reproductive Biology
  - Contribución Rocío Aznar Gimeno: Conceptualización, metodología, software, validación, análisis formal, investigación, visualización, preparación, revisión, edición y escritura del manuscrito.

- Gargallo-Puyuelo, C. J., Aznar-Gimeno, R., Carrera-Lasfuentes, P., Lanas, A., Ferrandez, A., Quintero, E., Carrillo, M., Alonso-Abreu, I., Esteban L. M., de

la Vega Rodrigálvarez-Chamarro M., Del Hoyo-Alonso, R., & García-González, M. A. (2022). Predictive Value of Genetic Risk Scores in the Development of Colorectal Adenomas. Digestive Diseases and Sciences, 67(8), 4049-4058.

- Factor de impacto JCR 2021-2022: 3,487 - 3,1. 57/93 (Q3)
- Categoría: Gastroenterology & Hepatology
- Contribución Rocío Aznar Gimeno: Metodología, software, validación, análisis formal, investigación, curación de datos, visualización, preparación del borrador original del manuscrito, revisión, edición y escritura del manuscrito final. Carla J. Gargallo-Puyuelo y Rocío Aznar-Gimeno comparten primera autoría.

- Aznar-Gimeno, R., Carrera-Lasfuentes, P., del-Hoyo-Alonso, R., Doblaré, M., & Lanas, Á. (2021). Evidence-based selection on the appropriate FIT cut-off point in CRC screening programs in the COVID pandemic. Frontiers in medicine, 8, 712040.

  - Factor de impacto JCR 2021: 5,058. 53/172 (Q2)
  - Categoría: Medicine, General & Internal
  - Contribución Rocío Aznar Gimeno: Metodología, software, validación, análisis formal, investigación, curación de datos, visualización, preparación del borrador original del manuscrito, revisión, edición y escritura del manuscrito final.