REV CUBANA MED TROP 2012;64(1):35-42

Clasificación de dengue hemorrágico utilizando árboles de decisión en la fase temprana de la enfermedad

MSc. Beatriz Vega Riverón,[1] Dr. C. Lizet Sánchez Valdés,[1] Dr. C. José Cortiñas Abrahantes,[1] MSc. Osvaldo Castro Peraza,[1] Dr. C. Daniel González Rubio,[1] MSc. Marta Castro Peraza[1]

RESUMEN

Introducción: el dengue es una enfermedad viral con comportamiento epidémico, a su inicio no es posible saber qué pacientes evolucionarán desfavorablemente, sin embargo, pueden presentar signos de alarma que anuncian deterioro clínico. **Objetivo**: aplicar la técnica de árboles de decisión a la búsqueda de signos de alarma de gravedad en la fase temprana de la enfermedad. **Métodos**: la muestra de estudio la constituyeron 230 pacientes ingresados con dengue en el Instituto de Medicina Tropical "Pedro Kourí" en 2001. Las variables consideradas para la clasificación fueron los signos, síntomas y exámenes de laboratorio al tercer día de evolución de la enfermedad. Se aplicó el algoritmo de árboles de clasificación y regresión utilizando el índice de Gini. Se consideraron diferentes matrices de pérdida para mejorar la sensibilidad. **Resultados**: el algoritmo ARC, correspondiente a la mejor pérdida, tuvo una sensibilidad de 98,68 % y error global de 0,36. Sin considerar pérdida, el árbol resultante obtuvo una sensibilidad de 74 % con un error de 0,25. En ambos casos las variables de mayor importancia fueron plaqueta y hemoglobina. **Conclusiones**: se proponen reglas de decisión con alta sensibilidad y valor predictivo negativo de utilidad en la práctica clínica. Las variables de laboratorio resultan tener mayor importancia que las clínicas para discriminar las formas clínicas de dengue.

Palabras clave: fiebre hemorrágica del dengue, dengue hemorrágico, árboles de decisión, predicción, Cuba.

INTRODUCCIÓN

El dengue es una enfermedad viral con tendencia epidémica, transmitida al hombre por la picada de mosquitos del género *Aedes*.[1] Se estima que anualmente ocurren entre 50 y 100 millones de infecciones y 36 millones son casos sintomáticos, de los cuales más de 2 millones desarrollan formas graves.[2]

La infección por dengue se manifiesta con un amplio espectro que incluye desde los infectados asintomáticos hasta casos graves. Se han descrito fundamentalmente 2 formas clínicas: la fiebre del dengue (FD) o dengue clásico y la fiebre hemorrágica del dengue/síndrome de choque por dengue (FHD/SCD).[3]

Al inicio de la enfermedad no es posible conocer qué pacientes tendrán complicaciones y evolucionarán a FHD/SCD. Sin embargo, pueden presentar manifestaciones clínicas que anuncien el deterioro cuando aún su cuadro clínico no cumple los criterios para clasificarlo como caso de FHD/SCD. Son los llamados signos de alarma (SA), cuya identificación en los días u horas previas al choque es fundamental para establecer una correcta intervención terapéutica temprana que, hasta el momento, es la medida más eficaz para disminuir la probabilidad de muerte del paciente. Por esa razón es que cada vez cobran mayor importancia los estudios y el desarrollo de métodos con el propósito de encontrar un conjunto de signos tempranos que tengan valor predictivo positivo con respecto al desarrollo de dengue hemorrágico.[4-6]

Los métodos de clasificación son herramientas adecuadas para abordar este problema y pueden servir con 2 propósitos. El primero cuando no se conocen las clases de antemano (no se conoce

[1] Instituto de Medicina Tropical "Pedro Kourí". La Habana, Cuba.

el conjunto de signos predictivos) y se quiere encontrar conglomerados dentro del conjunto de observaciones (aprendizaje *no supervisado)*. El segundo, cuando se conocen *a priori* las diferentes clases y el objetivo es establecer una regla que permita clasificar una nueva observación en una de las clases ya existentes (*aprendizaje supervisado*), a partir de un conjunto de datos correctamente clasificados (conocido como conjunto de entrenamiento).[7]

Este trabajo se enfoca en la aplicación de la técnica clasificatoria de árboles de regresión y clasificación (ARC), para hallar reglas de decisión que permitan clasificar un paciente con dengue en las diversas formas de la enfermedad a partir de características clínicas y de laboratorio.[8] El desempeño se evaluó sobre la base de la capacidad del método de reducir la tasa de error global y su habilidad de clasificar correctamente a los pacientes con FHD/SCD.

MÉTODOS

Selección de la muestra: estuvo constituida por 230 pacientes del total ingresado por dengue producido por el serotipo III, en el Instituto de Medicina Tropical "Pedro Kourí" en el período de junio de 2001 a marzo de 2002 y en los cuales fue demostrada la infección por dengue. La confirmación de la enfermedad se realizó mediante las pruebas serológicas IgM o IgG específica.

La muestra incluyó a los 76 pacientes que presentaron FHD/SCD. Para seleccionar los casos con dengue clásico se calculó un tamaño de muestra de 154 pacientes, al estimarse una prevalencia de 20 % del factor de exposición en los no enfermos para 95 % de confiabilidad.

DEFINICIÓN DE CASO FHD/SCD

Se consideró como FHD/SCD a todo paciente con diagnóstico de infección por dengue y que además reuniera los criterios que recomienda la OMS/OPS[9] hasta 2009, que incluyen: (I) fiebre, (II) alguna manifestación hemorrágica, (III) trombocitopenia (plaquetas $\leq 100 \times 10^9$/L) y (IV) evidencias de extravasación de plasma.

Recolección y procesamiento de los datos: las variables consideradas para la clasificación fueron la presencia o no de los síntomas siguientes: fiebre (F), cefalea (C), dolor retroocular (DO), mialgias (M), vómitos (V), dolor abdominal (DA), altralgias (A), diarreas (D), *rash* (R), petequias (P), gingivorragia (G), prurito (P), epistaxis (E), astenia (AS), hematemesis (HM), lipotimia (LI); y las variables de laboratorio: leucocituria (LEU), hematuria (HR), cilindruria (CI), hemoglobina (HB), hematocrito (HTO), leucocitos (LTO), plaquetas (PLQ), transaminasa glutámico oxalacético (TGO), transaminasa glutámico pirúvico (TGP), creatinina (CR). La información sobre las variables de interés se recolectó a partir de la revisión de las historias clínicas.

Se conoce que los primeros signos de alarma se presentan durante la fase inicial de la enfermedad, que suele durar alrededor de 3 d y en la fase crítica que transcurre entre el cuarto y séptimo día.[9] Para poder clasificar a los pacientes tempranamente, se utilizó como punto de corte la primera medición de las variables antes mencionadas que se le hizo al paciente a partir del tercer día de evolución de la enfermedad.

Análisis estadístico: con propósitos descriptivos se utilizaron diagramas de barra y de caja. Para encontrar reglas de decisión que permitieran clasificar a los pacientes en FD y FHD/SCD se aplicó el algoritmo de árboles de clasificación y regresión propuesto por *Breiman* y otros en1984.[8] Este es un método en el cual, siguiendo reglas de ramificación específica, se dividen los datos en subconjuntos mutuamente excluyentes. Este proceso es repetido varias veces dentro de cada subconjunto tratando de minimizar el error de clasificación. Por último se obtiene un árbol en el que cada rama es una regla de decisión. Como medida del error de clasificación se utilizó el índice de Gini. La importancia de las variables se midió a través de la disminución que produce esa variable en el error de cada uno de los nodos del árbol final.[7]

En el dengue, las consecuencias de clasificar erróneamente a un paciente que es un caso de FHD/SCD como FD pueden ser fatales, por lo que en el proceso de clasificación se le debe prestar especial interés a este tipo de error. Eso se puede tener en cuenta incorporando en la modelación determinadas matrices de pérdida y modificando

el índice de Gini.[7] Se consideraron las matrices de pérdidas siguientes:

$$\begin{pmatrix} 0 & 1{,}001\text{-}i/100 \\ i/100 & 0 \end{pmatrix} \quad \text{donde } i = 1,...,100$$

Con esto lo que se persigue es alterar la probabilidad *a priori* que tiene un paciente de tener una forma clínica u otra, dándole mayor peso a la clasificación correcta de los pacientes con FHD/SCD.

El método ARC se evaluó sobre la base de su capacidad de minimizar la tasa de error global. En el caso en que no se considera pérdida (no se hizo distinción entre un caso erróneamente clasificado de dengue clásico o de FHD/SCD) el método no asegura una sensibilidad alta, y en dependencia de la pérdida (diferentes pesos son considerados según el tipo de error de clasificación) se puede asegurar una sensibilidad alta o no.

Para estimar la tasa de error global se empleó la técnica de validación cruzada estratificada con 10 particiones. Consiste en dividir los datos de modo aleatorio en 10 subconjuntos mutuamente excluyentes. Luego repetir el proceso siguiente 10 veces: en la iteración i, el conjunto i-esimo es utilizado como conjunto de prueba para la predicción y los subconjuntos restantes (conjunto de entrenamiento) para obtener el clasificador.

Los métodos se implementaron en R versión 2.10, el paquete utilizado fue: RPART en donde se encuentran las funciones básicas para implementar el algoritmo ARC.

RESULTADOS

En la figura 1 se muestra el porcentaje de casos según formas clínicas del dengue para los diferentes síntomas recogidos. Se puede apreciar que en ambas formas los síntomas más frecuentes resultaron: fiebre, cefalea, mialgias, artralgias, y dolor retro-ocular que caracterizan el cuadro clínico del dengue. La astenia, los vómitos y la anorexia le siguen en frecuencia con un predominio de FHD/SCD.

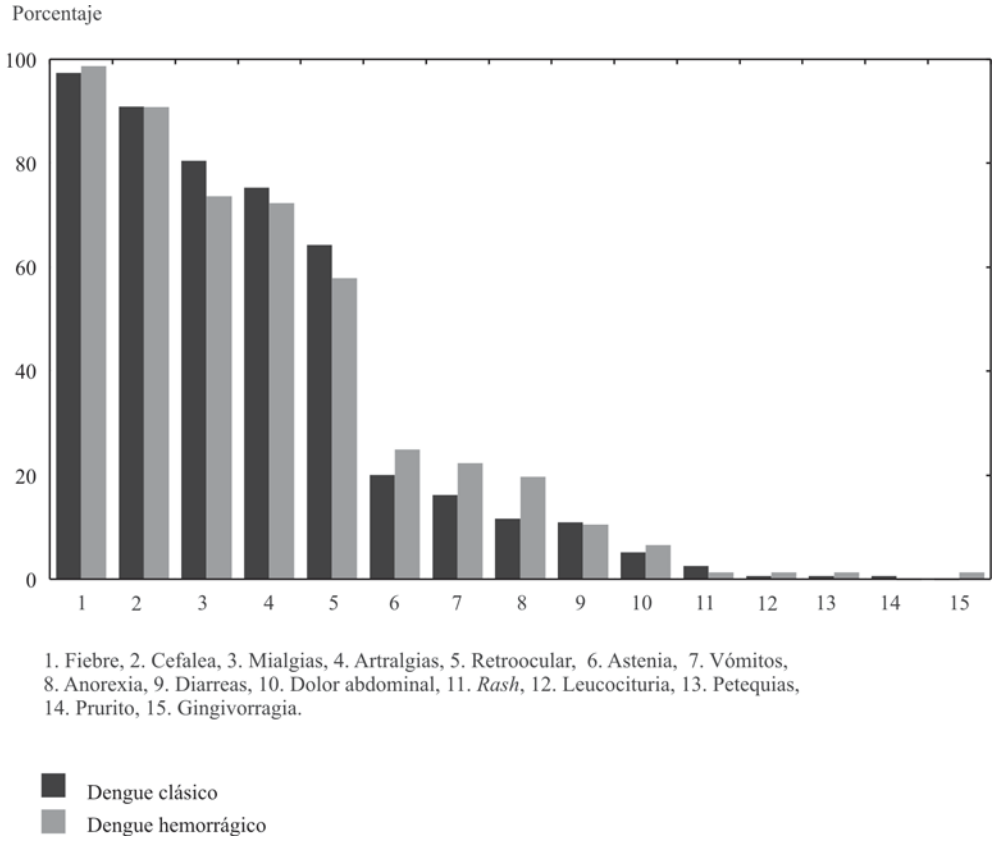

Fig. 1. Distribución de los síntomas según formas clínicas del dengue. Instituto "Pedro Kourí", 2001-2002.

En la figura 2 se muestra el comportamiento de las variables de laboratorio según las formas clínicas de la enfermedad. Se aprecia que los valores de plaqueta en la forma hemorrágica estuvieron por debajo de los valores de los pacientes con dengue clásico. En cambio, para el hematocrito y la hemoglobina se observó un ligero incremento en los casos con FHD/SCD. Para el resto de las variables la distribución en ambos grupos fue similar.

En la figura 3 se puede apreciar el árbol de clasificación final sin pérdida y un gráfico correspondiente a la importancia de las variables. La variable de mayor importancia resultó las plaquetas, en particular utilizando como punto de corte para estas un conteo de 65,5, se obtuvo que 44 pacientes (58,9 %) fueron correctamente clasificados como FHD/SCD (correspondiente a la rama conteo de plaqueta menor que 65,5) y 138 (89,6 %) como dengue clásico (correspondiente a la rama conteo de plaqueta mayor que 65,5). La segunda variable que desempeña un papel importante en la clasificación es la hemoglobina, que al menos es 35,7 % tan importante como las plaquetas. El hematocrito resultó la tercera variable más importante para clasificar a un paciente en las formas clínicas del dengue, aun cuando no fue utilizada en el proceso de ramificación del árbol. Utilizando las reglas de decisión (correspondiente a cada rama del árbol) para clasificar los pacientes de la muestra, se obtuvo que 19 (12,3 %) pacientes con dengue clásico y 20 (26,3 %) pacientes con FHD/SCD fueron erróneamente clasificados, lo que representa una tasa de error total de 16,9 %. Al aplicar el método de validación cruzada se obtuvo que de los 154 pacientes ARC, clasificaron de modo incorrecto 31 pacientes (20,1 %) y en los casos hemorrágicos 28 (36,8 %), que representa una tasa de error total de 25,6 %.

En la figura 4 se muestran los resultados correspondientes a cada una de las matrices de pérdidas consideradas. Se puede apreciar que la tasa de error global siempre toma valores superiores a 30 % (representado por la línea horizontal) e inferiores a 40 % para la mayoría de las matrices de

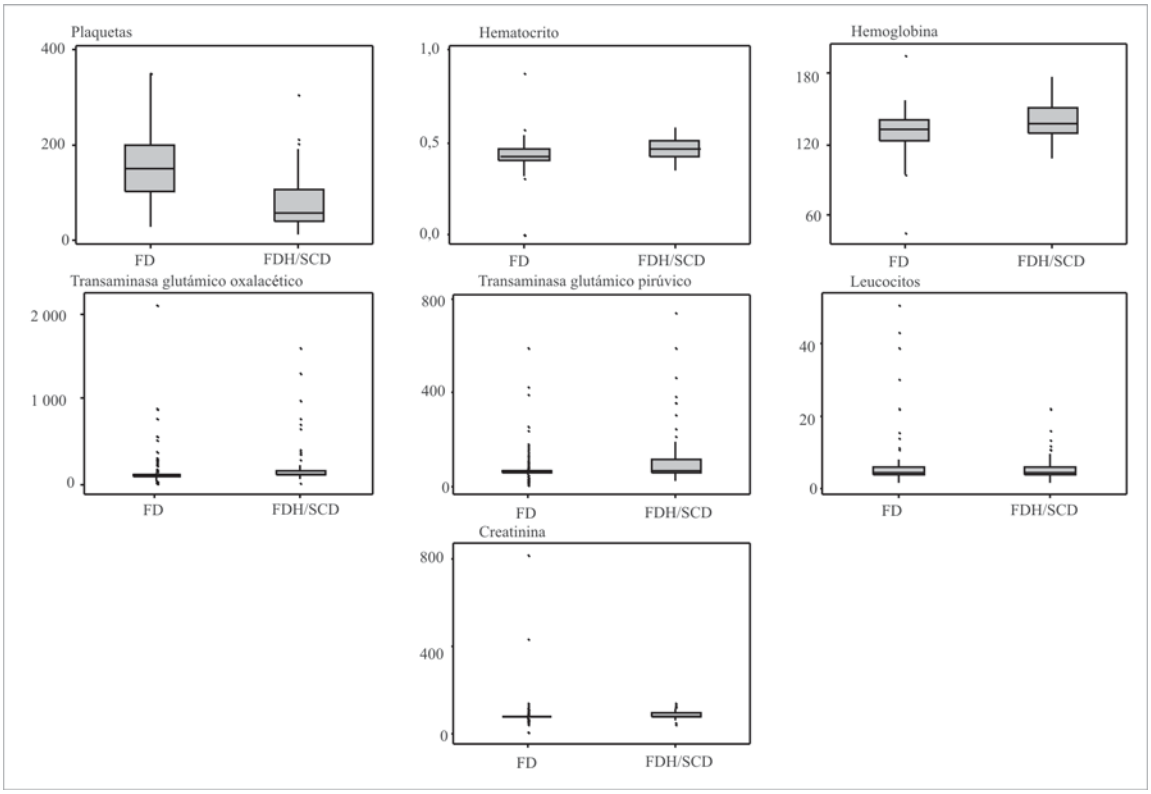

Fig. 2. Diagrama de caja de las variables de laboratorio según formas clínicas del dengue. Instituto "Pedro Kourí", 2001-2002.

pérdidas consideradas. En contraste, el porcentaje de pacientes con FHD/SCD clasificados de manera correcta fluctúa de modo decreciente desde 90 % a valores menores que 20 %. El valor mínimo de error global estimado fue de 26 %, similar al obtenido en el modelo ajustado sin considerar pérdidas. La matriz correspondiente fue:

$$\begin{pmatrix} 0{,}00 & 0{,}931 \\ 0{,}07 & 0{,}000 \end{pmatrix}$$

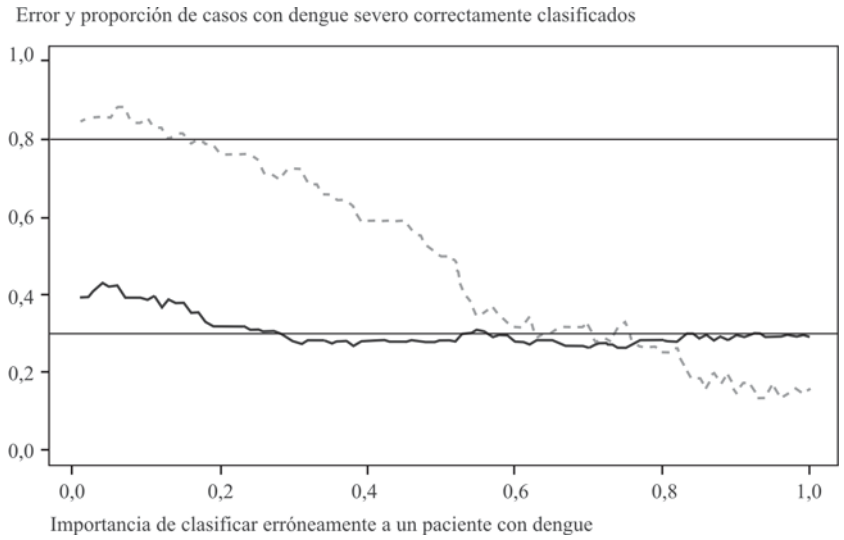

Fig. 4. Tasa de error global estimada (línea sólida) y proporción de pacientes con fiebre hemorrágica del dengue/síndrome de choque por dengue (FHD/SCD) correctamente clasificados (línea discontinua) para difertentes matrices de pérdida.

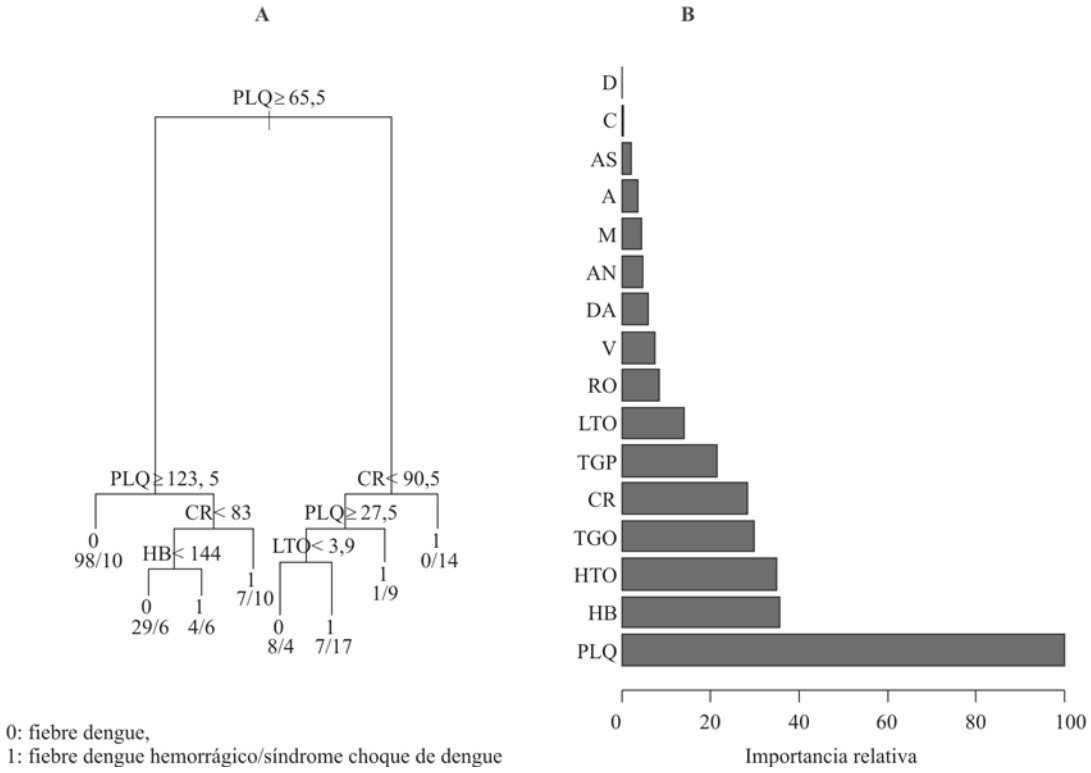

Fig. 3. A. Árbol de clasificación de los síntomas, signos y parámetros de laboratorio sin considerar pérdida. **B.** Gráfico de la importancia de las variables sin considerar pérdida.

En la figura 5 se muestran los resultados correspondiente a esta pérdida.

En este caso se obtuvo de nuevo que la variable más importante que puede ser utilizada para discriminar a los pacientes son las plaquetas y en segundo lugar la hemoglobina (Fig. 5B). Vale destacar que el árbol de clasificación obtenido incluyó la variable conteo de hematocritos (que en el caso anterior no estaba contemplada) (Fig. 5A). Como este fue el árbol obtenido considerando la mejor

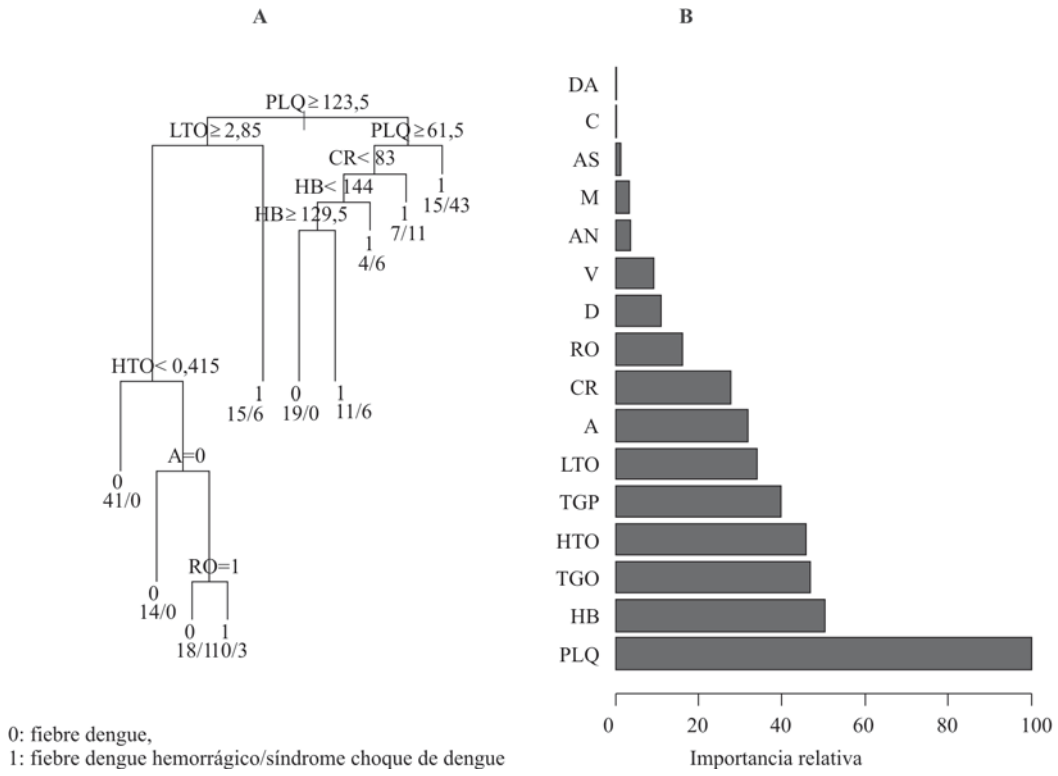

Fig 5. A. Árbol de clasificación de los síntomas, signos y parámetros de laboratorio considerando la mejor pérdida. **B.** Gráfico de la importancia de las variablesconsiderando la mejor pérdida.

Tabla. Indicadores de eficiencia

a) Conjunto de entrenamiento	Árboles de regresión y clasificación (ARC)	
	No pérdida	Mejor pérdida
Índice	Valor (IC 95 %)	Valor % (IC 95 %)
Sensibilidad	73,68 (63,13-84,24)	98,68 (95,46-100,00)
Especificidad	87,66 (82,14-93,18)	59,74 (51,67-67,81)
Índice de validez	83,04 (77,98-88,11)	72,61 (66,63 8,59)
Valor predictivo +	74,67 (64,16-85,18)	54,74 (46,04-63,44)
Valor predictivo -	87,10 (81,50-92,70)	98,92 96,29-100,00
b) Validación cruzada		
Sensibilidad	57,89 (46,14-69,65)	86,84 (78,58-95,10)
Especificidad	83,77 (77,62-89,92)	42,86 (34,72-51,00)
Índice de validez	75,22 (69,42-81,01)	57,39 (50,78-64,00)
Valor predictivo +	63,77 (51,70-75,83)	42,86 (34,72-51,00)
Valor predictivo -	80,12 (73,65-86,60)	86,84 (78,58-95,10)

pérdida posible, quiere decir que casi no se cometió ninguna mala clasificación en los pacientes FHD/SCD (1,3 %), pero sí en los casos con dengue clásico (52 %) para una tasa de error global de 36 %.

En la tabla 1 se muestran algunos indicadores estadísticos que se utilizan para evaluar la precisión de un método de clasificación. Se puede apreciar que la sensibilidad y el valor predictivo negativo correspondiente al análisis teniendo en cuenta la mejor pérdida, fue mayor que cuando no se consideró esta, tanto en el conjunto de entrenamiento (tabla 1a) como al utilizar validación cruzada (tabla 1b).

DISCUSIÓN

El estudio permitió identificar marcadores de laboratorio asociados a las formas clínicas del dengue. Todos los métodos identificaron como variable discriminatoria más importante el conteo de plaquetas, la cual en estudios clínicos realizados, aparece como un indicador de alarma.[10] Otro aporte importante son las reglas de decisión obtenidas por el algoritmo ART, las cuales pueden ser vistas como signos de alarma, y además servir como guías para el procedimiento clínico de los pacientes al tercer día de evolución de la enfermedad.

La baja estimación de los valores predictivos positivos a DHF/SCD cuando el método fue utilizado con nuevas observaciones, pudo deberse a que el estudio es realizado al tercer día de la enfermedad, momento para el cual el procedimiento clínico pudo haber influido significativamente en la evolución del paciente. Para obtener una estimación más consistente de estos indicadores se recomienda utilizar varios métodos de validación, aplicados en un conjunto de individuos que no hayan sido utilizados de ninguna forma en la obtención del clasificador que se está evaluando.[11] En este caso no fue posible seguir esa estrategia debido a la poca disponibilidad de datos.

Existen varios reportes de estudios relacionados con dengue en los que se ha empleado esta metodología. *Lee* y otros[4] emplearon árboles de decisión con el objetivo de decidir si se hospitaliza o no a un paciente sospechoso de dengue. En el análisis se incluyeron no solo variables clínicas y de laboratorio (como en nuestro estudio), sino también variables demográficas como el sexo y la edad. El árbol final tuvo una sensibilidad de 100 % pero una especificidad de 46 %, similar a lo obtenido en este estudio para el caso en que se aplicó CART teniendo en cuenta la mejor pérdida. Otros autores como *Tanner* y otros[12] incluyeron en el análisis datos de naturaleza virológica y clínica como el conteo de linfocitos. En este caso el árbol obtenido incluyó como primera variable discriminatoria el conteo de plaquetas (siendo utilizada varias veces como variable de ramificación), la creatinina y al final del árbol el hematocrito, resultado que coincide con lo obtenido en este trabajo.

Esta metodología no solo ha mostrado buenos resultados en estudios en adultos sino también en estudios pediátricos como muestran *James* y otros.[13] En la cual se utilizó un árbol de clasificación y regresión para obtener un algoritmo predictivo de la forma severa. El algoritmo identificó como indicadores importantes el conteo de plaquetas, hematocrito, conteo de células blancas y la edad.

Se han realizado estudios previos de este mismo grupo de individuos y las manifestaciones que predominaron la forma clínica fueron el vómito y el dolor abdominal, variables que no resultaron tener importancia en este estudio. Eso pudo deberse a la diferencia de los puntos de cortes para las mediciones entre los estudios.[3]

El método ARC puede constituir un método práctico para proporcionar reglas de decisión que logran ser útiles en la práctica clínica y en la elaboración de sistemas de expertos para clasificar a un paciente. Este estudio tiene en cuenta solo la evolución de la enfermedad a partir del tercer día de comenzada la fiebre, sin embargo, sería interesante, en estudios futuros emplear métodos para datos longitudinales que permitan incorporar la dinámica de la enfermedad en el tiempo.

Classification of dengue hemorrhagic fever using decision trees in the early phase of the disease

ABSTRACT

Introduction: dengue is a viral disease with endemic behavior. At the beginning of the illness it is not possible to know which patients will have an unfavorable evolution and develop a severe

form of dengue. However, some warning symptoms and signs may be present. **Objective**: to apply decision tree techniques to the exploration of signs of severity in the early phase of the illness. **Methods**: the study sample was made up of 230 patients admitted with dengue to "Pedro Kourí" Institute of Tropical Medicine in 2001. The variables considered for the classification were the signs, symptoms and laboratory exams on the third day of evolution of the illness. The algorithm of classification and regression trees using the Gini's index was applied. Different loss matrices to improve the sensitivity were considered. **Results**: the algorithm CART, corresponding to the best loss, had a sensitivity of 98,68% and global error of 0,36. Without considering loss, it obtained its sensitivity reached 74% with an error of 0,25. In both cases, the most important variables were platelets and hemoglobin. **Conclusions**: the study submitted rules of decision with high sensitivity and negative predictive value of utility in the clinical practice. The laboratory variables resulted more important from the informational viewpoint than the clinical ones to discriminate clinical forms of dengue.

Key words: dengue hemorrhagic fever, hemorrhagic dengue, tree decision, prediction, Cuba.

REFERENCIAS BIBLIOGRÁFICAS

1. Guzmán MG, García G, Kourí G. El dengue y el dengue hemorrágico: prioridades de investigación. Rev Panam Salud Publica. 2006;19:204-15.
2. Díaz FA, Martínez RA, Villar LA. Criterios clínicos para diagnosticar el dengue en los primeros días de enfermedad. Biomédica. 2006;26:22-30.
3. González D, Castro O, Rodríguez F, Portela D, Gracés M, Martínez A, et al. Descripción de la fiebre hemorrágica del dengue, serotipo 3, Ciudad de La Habana, 2001-2002. Rev Cubana Med Trop. 2008;60(1):48-54.
4. Lee VJ, Lye DC, Sun Y, Leo YS. Decision tree algorithm in deciding hospitalization for adult patients with dengue haemorrhagic fever in Singapore. Trop Med Int Health. 2009;I4:II54-9.
5. Martínez E. Dengue y dengue hemorrágico: aspectos clínicos. Salud Pública Mex. 1995;37:S29-S44.
6. Lee VJ, Lye DC, Sun Y, Fernandez G, Ong A, Leo YS. Predictive value of simple clinical and laboratory variables for dengue hemorrhagic fever in adults. J Clin Virol. 2008;42:34-9.
7. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Data Mining, inference and predictions. New York: Springer-Verlag; 2001.
8. Mackie I. M B. Principles of Data Mining. London: Springer-Verlag; 2007.
9. PAHO. Dengue and dengue hemorrhagic fever in the Americas guidelines for prevention and control. Washington: PAHO; 1994. (Scientific Publication No. 548)
10. Castro OE, González D, Pelegrino JL, Guzmán MG, Kourí G. Dengue y dengue hemorrágico en Cuba- Aportes a la clínica y manejo de casos. Rev Panam Infectol. 2004;6:39-42.
11. Abrahantes JC, Aerts M, Everbroeck BV, Saegerman C, Berkvens D, Geys H, et al. Classification of sporadic Creutzfeldt-Jakob disease based on clinical and neuropathological characteristics. Eur J Epidemiol. 2007;22:457-65.
12. Tanner L, Schreiber M, Low JG, Ong A, Tolfvenstam T, Lai YL, et al. Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness. PloS Negl Trop Dis. 2008;2:1-9.
13. Potts JA, Gibbons RV, Rothman AL, Srikiatkhachorn A, Thomas SJ, Supradish P, et al. Prediction of dengue disease severity among pediatric thai patients using early clinical laboratory indicators. PloS Negl Trop Dis. 2010;4:769-76.

Recibido: 28 de marzo de 2011. Aprobado: 13 de septiembre de 2011.
Beatriz Vega Riverón. Instituto de Medicina Tropical "Pedro Kourí". Autopista Novia del Mediodía, Km 6 ½. Lisa, La Habana, Cuba. CP 17100. Correos electrónicos: beatrizv@ipk.sld.cu, lsanchez@ipk.sld.cu, daniel@ipk.sld.cu, josecortinas@hotmail.com, osvaldo@ipk.sld.cu, martac@ipk.sld.cu