

ARTÍCULO ORIGINAL

Predicción de la epidemia del virus sincitial respiratorio en Bogotá, D.C., utilizando variables climatológicas

Gilberto González-Parra^{1,2}, José F. Querales¹, Diego Aranda³

¹ Grupo de Matemática Multidisciplinar, Universidad de Los Andes, Mérida, Venezuela

² Centro de Investigaciones en Matemática Aplicada, Universidad de Los Andes, Mérida, Venezuela

³ Grupo Signos, Departamento de Matemáticas, Facultad de Ciencias, Universidad El Bosque, Bogotá, D.C., Colombia

Introducción. El virus sincitial respiratorio es uno de los principales causantes de mortalidad de niños y adultos mayores en el mundo.

Objetivo. Predecir las semanas de inicio de los brotes de infección por el virus sincitial respiratorio en Bogotá entre 2005 y 2010, utilizando variables climatológicas como variables de predicción.

Materiales y métodos. Se establecieron las fechas de inicio de los brotes en niños menores de cinco años ocurridas entre 2005 y 2010, en Bogotá D.C., Colombia. Se seleccionaron las variables climatológicas utilizando una matriz de correlación y, posteriormente, se construyeron 1.020 modelos resultado de la combinación de las distintas variables climatológicas y las semanas de antelación al inicio del brote. Se seleccionaron los modelos utilizando los datos correspondientes a periodos de seis (2005-2010), cuatro (2005-2008) y dos años (2009-2010). Se determinaron los mejores modelos y las variables climatológicas más relevantes, utilizando clasificadores bayesianos ingenuos y curvas características de operación del receptor (*Receiver Operating Characteristic*, ROC).

Resultados. Los mejores resultados se obtuvieron con los modelos que utilizaron el periodo de dos años (2009-2010) y la semana 0, con 52 y 60 % de aciertos, respectivamente. La humedad mínima fue la variable que más apareció en los mejores modelos (62 %). Los clasificadores bayesianos ingenuos permitieron establecer cuáles eran los mejores modelos para predecir la semana de inicio del brote.

Conclusiones. Los resultados sugieren que los modelos en que se utilizaron la humedad mínima, la velocidad del viento y la temperatura mínima serían los modelos de predicción más eficaces.

Palabras clave: teorema de Bayes, virus sincitiales respiratorios, predicción, climatología, epidemias.
doi: <http://dx.doi.org/10.7705/biomedica.v36i3.2763>

Prediction of the respiratory syncytial virus epidemic using climate variables in Bogotá, D.C.

Introduction: The respiratory syncytial virus is one of the most common causes of mortality in children and older adults in the world.

Objective: To predict the initial week of outbreaks and to establish the most relevant climate variables using naive Bayes classifiers and receiver operating characteristic curves (ROC).

Materials and methods: The initial dates of the outbreaks in children less than five years old for the period 2005-2010 were obtained for Bogotá, Colombia. We selected the climatological variables using a correlation matrix and we constructed 1,020 models using different climatological variables and data from different weeks previous to the initial outbreak. In addition, we selected models using a six-year period (2005-2010), a four-year period (2005-2008), and a two-year period (2009-2010). We obtained the best predictive models and the most relevant climatological variables to predict the outbreak using naive Bayes classifiers and ROC curves.

Results: The best models were those using a two-year period (2009-2010) and week 0, with 52% and 60% of effectiveness, respectively. Humidity was the most frequent variable in the best models (62%).

Conclusions: We used naive Bayes classifiers to establish the best models to predict correctly the initial week of the outbreak. Our results suggest that the best models used humidity, wind speed and minimum temperature in outbreaks prediction.

Key words: Bayes theorem, respiratory syncytial viruses, forecasting, climatology, epidemics.
doi: <http://dx.doi.org/10.7705/biomedica.v36i3.2763>

Contribución de los autores:

Gilberto González-Parra: diseño del estudio y dirección de la investigación

José F. Querales: depuración de los datos, cálculos de las curvas ROC y de probabilidades

Diego Aranda: diseño del trabajo, recolección de datos

Todos los autores participaron en el análisis de los datos y la escritura del manuscrito.

El virus sincitial respiratorio (RSV) fue aislado por primera vez en 1956 por Morris, *et al.*, en un grupo de chimpancés que padecían coriza, por lo que los autores lo denominaron agente de la coriza en chimpancés. Al año siguiente, Chanock y su equipo aislaron el mismo agente en dos niños, uno con diagnóstico de neumonía y el otro con laringitis estridente (espasmo laríngeo o *croup*); desde ese momento, se lo denominó con el nombre actual, en razón de su efecto citopático característico (1).

Actualmente, una de las principales causas de mortalidad de niños y adultos mayores en el mundo, especialmente en países en desarrollo, son las infecciones agudas de las vías aéreas inferiores, las cuales provocan, aproximadamente, un tercio de todas las muertes estimadas en niños, menores de cinco años (1-3).

Se han propuesto diferentes modelos matemáticos epidemiológicos para el análisis de diversas enfermedades, entre ellas, la infección por el virus sincitial respiratorio (4-8). También, se han utilizado modelos basados en ecuaciones diferenciales para explicar la dinámica del RSV en la población (4,5,7). En uno de los estudios se hizo una descripción retrospectiva del virus de la influenza A (H1N1) a nivel poblacional y se concluyó que durante el 2009 hubo un aumento significativo de casos, especialmente en los meses de agosto y septiembre (6). En otro trabajo, se desarrolló un modelo basado en redes o grafos para describir la evolución del número de casos de personas infectadas con el virus A (H1N1) (8).

En Bogotá se han reportado muestras positivas para RSV de los tipos A y B, y en algunos casos, se registró la infección simultánea con los dos tipos (9). Por otra parte, en 2009, en los informes de vigilancia centinela se registró la presencia del virus sincitial respiratorio desde las primeras semanas epidemiológicas del año. A partir de la semana 10, su proporción aumentó hasta alcanzar 60 % de las muestras con resultado positivo, porcentaje que disminuyó ligeramente hacia la semana 22, lo cual coincidió con lo registrado en años anteriores (6).

Correspondencia:

Gilberto González-Parra, Grupo de Matemática Multidisciplinar, Facultad de Ingeniería, Núcleo Universitario "Pedro Rincón Gutiérrez", La Hechicera, edificio C, Universidad de Los Andes, Mérida, Venezuela
Teléfono: (001-817) 403 9588
gcarlos@ula.ve

Recibido: 23/04/15; aceptado: 29/03/16

La epidemiología ambiental es parte de un nuevo paradigma basado en un enfoque sistémico y multidisciplinario para el estudio de parámetros ambientales en regiones de transmisión de enfermedades. Constituye una innovadora línea de trabajo que permite la caracterización a gran escala de factores que condicionan la reproducción, el desarrollo y el mantenimiento de un virus, mediante el análisis de los datos revelados por los satélites y otros equipos. El principal objetivo de este trabajo fue predecir el inicio de un brote epidémico del virus sincitial respiratorio, utilizando variables climatológicas.

Los clasificadores bayesianos se han utilizado en diversos campos, como la inteligencia artificial y el diagnóstico médico (Larrañaga P. Learning bayesian networks from data. Some applications in biomedicine. IDAMAP 2002: Intelligent Data Analysis in Medicine and Pharmacology. Lyon, July 23rd, 2002). Dada la gran cantidad de datos meteorológicos disponibles actualmente, y la correlación entre estos y las epidemias respiratorias, se propuso la construcción de modelos con clasificadores bayesianos ingenuos para predecir el inicio de los brotes de infección por RSV en Bogotá. En un trabajo similar con datos del condado de Salt Lake City en Utah, Estados Unidos (10), se encontró que la temperatura y la velocidad del viento eran los mejores predictores y se logró pronosticar la epidemia hasta con tres semanas de antelación.

Materiales y métodos

Los clasificadores bayesianos ingenuos son un tipo de grafos o redes que han sido de gran ayuda para resolver problemas en el área de la medicina. Una red bayesiana es un grafo acíclico dirigido, en el cual cada nodo representa una variable y, cada arco, una dependencia probabilística. La variable a la que apunta el arco es dependiente (en términos de causa y efecto) de la que está en su origen. En cada variable se especifica la probabilidad condicional con respecto a las variables padre, en caso de haberlas. Una red bayesiana tiene, al menos, una raíz (nodo sin variable padre) y un nodo terminal (nodo sin 'hijos'). En una red bayesiana cualquiera, el cálculo de la probabilidad conjunta de las n variables se puede representar como:

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i | X_{j(i)}),$$

donde $\{X_{1(i)}, X_{2(i)}, \dots, X_{m(i)}\}$ es el conjunto de variables padre de las X_i .

La red bayesiana puede construirse valiéndose exclusivamente del conocimiento experto; sin embargo, se han publicado diversos métodos para el aprendizaje con base en un conjunto de datos (Larrañaga P. Learning bayesian networks from data. Some applications in biomedicine. IDAMAP 2002: Intelligent Data Analysis in Medicine and Pharmacology. Lyon, July 23rd, 2002). En investigaciones recientes, se han propuesto métodos que combinan ambos enfoques (Steck H, Tresp V. Bayesian belief networks for data mining. Proceedings, Second Workshop on Data Mining und Data Warehousing als Grundlage Moderner Entscheidungsunterstuezender Systeme, DWDW99, Sammelband, Universität Magdeburg, September, 1999). En este estudio, la red bayesiana se construyó con los valores de las variables climatológicas y los datos sobre la aparición del brote epidémico de la infección por el RSV.

El clasificador bayesiano ingenuo (11) es un modelo de clasificación construido bajo la premisa de que todas las variables predictoras son condicionalmente independientes dada la variable de clase (C). Sin embargo, cuando se trata de problemas reales, muchas veces esta suposición no se cumple. En algunas ocasiones, no obstante, el paradigma bayesiano ingenuo ha demostrado un buen funcionamiento (12,13). El clasificador bayesiano ingenuo aplica el teorema de Bayes con el fin de predecir para cada instancia x la clase $c \in C$, con la máxima probabilidad *a posteriori*. En este caso particular, se utilizó la librería e1071 desarrollada por el *R Development Core Team*, la cual contiene la función bayesiana ingenua con la que se calcula el clasificador bayesiano ingenuo.

Otra herramienta utilizada en este trabajo fue la curva característica de funcionamiento del receptor (*Receiver Operating Characteristic*, ROC), la cual permite visualizar, organizar y seleccionar clasificadores o marcadores con base en su capacidad de discriminación en todo el rango de puntos de corte. Para evaluar la habilidad discriminatoria de un marcador, es común resumir la información de la curva ROC en un único valor global o índice (14). En este trabajo se usaron solo dos clases: cada objeto, sujeto o caso fue etiquetado con uno de los elementos del conjunto positivo (S) o negativo (N), prediciendo la clase a la que pertenecía el caso.

La aplicación de un marcador en situaciones como la descrita produce cuatro posibles resultados:

1) si el caso es positivo y se clasifica como positivo, se cuenta como un verdadero positivo;

2) si el caso es positivo y se clasifica como negativo, se cuenta como un falso negativo;

3) si el caso es negativo y se clasifica como negativo, se cuenta como un verdadero negativo, y

4) si el caso es negativo y se clasifica como positivo, se cuenta como un falso positivo.

Dado el marcador y el conjunto de sujetos, se pudo construir una matriz de confusión o tabla de contingencia 2 x 2 para representar la disposición de dicho conjunto (cuadro 1). Los números a lo largo de la diagonal principal de la matriz de confusión en el cuadro 1 representan las clasificaciones correctas y, los que están a lo largo de la diagonal secundaria, los errores (la confusión) entre las clases. Esta matriz fue el soporte para varios indicadores comúnmente utilizados en sistemas de diagnóstico (14), de la siguiente manera:

- fracción de verdaderos positivos (TPF), denominada también tasa de aciertos;

- fracción de falsos positivos (FPF), denominada también tasa de falsa alarma,

$$FPF = \frac{\text{Negativos clasificados incorrectamente (falsos positivos)}}{\text{Total de negativos}}$$

- precisión

$$\text{Precisión} = \frac{\text{Verdaderos positivos}}{\text{Verdaderos positivos} + \text{falsos positivos}}$$

- exactitud

$$\text{Exactitud} = \frac{\text{Verdaderos positivos} + \text{verdaderos negativos}}{\text{Positivos} + \text{negativos}}$$

Las curvas ROC son gráficos bidimensionales en los que la fracción de verdaderos positivos (TPF) es representada en el eje Y y, la fracción de falsos positivos (FPF), en el eje X, y muestran las compensaciones relativas entre los beneficios (verdaderos positivos) y los costos (falsos positivos).

Datos climatológicos

Los datos meteorológicos de Bogotá entre 2005 y 2010 provenían de la estación meteorológica del Aeropuerto El Dorado, y se obtuvieron en el sitio de

Cuadro 1. Matriz de confusión

Clases hipotéticas	Clases verdaderas	
	P	N
Sí	Verdaderos positivos	Falsos positivos
No	Falsos negativos	Verdaderos negativos

internet de la *National Oceanic and Atmospheric Administration* (NOAA) de Estados Unidos (*National Climatic Data Center*- <http://cdo.ncdc.noaa.gov/CDO/CDO>).

Selección de las variables climatológicas

El clasificador bayesiano ingenuo se utilizó en este trabajo para seleccionar las variables climatológicas que mejor predecían la semana de inicio del brote de la infección por RSV. Para seleccionar las variables climatológicas se revisaron varios trabajos de otros autores, en los cuales se tomaron en cuenta diversas variables climatológicas (15-17). Se seleccionaron las investigaciones que reportaban un efecto positivo o negativo en la relación entre la actividad del RSV y las condiciones meteorológicas.

En varias publicaciones se observó que las variables meteorológicas con mejor correlación con el SRV fueron la humedad, el punto de rocío, la temperatura, la sensación térmica, la velocidad del viento, la presión atmosférica, la precipitación y la exposición a la luz ultravioleta. En este trabajo, se descartó esta última como variable potencial porque no se encontraron suficientes datos para validar su inclusión en el modelo de predicción. Se descartó el punto de rocío y de sensación térmica, debido a que estas variables podían derivarse de otras ya incluidas en el modelo (temperatura y humedad, y velocidad del viento y temperatura, respectivamente). Las variables consideradas en los modelos del presente trabajo fueron las siguientes: la humedad, de la cual se tomaron dos valores, la relativa mínima y la máxima; la temperatura, de la cual se hicieron tres mediciones, la media diaria, la mínima diaria y la máxima diaria; la presión atmosférica; la velocidad del viento, y la precipitación. Por último, es importante mencionar que cada una de estas ocho variables se representó mediante una única medición diaria.

Clasificador bayesiano ingenuo basado en las variables climatológicas

Para analizar la efectividad de los clasificadores bayesianos ingenuos, se disponía de los datos climatológicos y sobre el RSV en Bogotá, correspondientes a seis años (2005 a 2010). Se hicieron tres estudios por separado con los datos de dos años (2009 y 2010), de cuatro años (2005 a 2008) y de seis años (2005 a 2010), para determinar qué modelos arrojaban la mejor predicción del brote de la infección por SRV, usando la ROC (18). La cantidad de posibles modelos basados en los

clasificadores bayesianos ingenuos se calculó tomando todas las combinaciones posibles de las ocho variables climatológicas seleccionadas, independientemente de su orden, por lo cual se generaron 255 clasificadores.

La efectividad de esos 255 clasificadores se calculó usando el programa estadístico del *R Development CoreTeam* en su versión 2.15.0 y el paquete *e1701* (19).

Los clasificadores bayesianos ingenuos se formaron con los valores de las variables meteorológicas registrados por días. El conjunto de datos meteorológicos se tomó en forma regresiva a partir del domingo siguiente al inicio del brote de la infección por RSV. Como se puede observar en la figura 1 A, para el modelo de la semana 0, o de inicio del brote, se tomaron los datos correspondientes a seis semanas (42 días antes del domingo siguiente a la semana de inicio del brote), como los datos de entrada.

Para establecer dichas semanas, se siguieron los siguientes pasos: se usaron cuatro semanas (28 días antes del domingo de inicio del brote) para la clase S, es decir, la semana del brote, se usaron tres semanas (21 días, de los cuales los primeros 14 fueron los últimos utilizados en la clase S y los otros siete, los de la semana anterior) para la clase N, y por último, se utilizaron tres semanas (21 días, de los cuales los primeros 14 fueron los mismos empleados en la clase N y, los últimos siete, los de la semana anterior) para la clase N, es decir, 28 valores para la clase S y 42 para la clase N.

Los clasificadores (modelos) se construyeron independientemente para predecir el inicio del brote en la misma semana (semana 0), una semana antes del brote (semana 1), la segunda semana previa a este (semana 2) y la tercera semana previa (semana 3).

Para la construcción de los otros modelos de las semanas 1, 2 y 3, se procedió de forma análoga a la del modelo de la semana 0, pero tomando en cuenta los datos de entrada de la semana del brote, es decir, una semana antes del inicio del brote para la semana 1 (figura 1 B), dos semanas antes de la semana del brote para la semana 2 (figura 1 C) y, por último, tres semanas antes del inicio del brote para la semana 3 (figura 1 D). Los valores para cada una de las variables climatológicas en las distintas temporadas consideradas fueron los siguientes: para los modelos que utilizaron los datos de seis años se emplearon 168 valores (6

A. Datos de entrada (días de la semana)									
F-8	F-7	F-6	F-5	F-4	F-3	F-2	F-1	Brote semana (0)	Datos de salida
					Semanas del brote				S
			Semanas antes del brote						N
		Dos semanas antes del brote							N

B. Datos de entrada (días de la semana)									
F-8	F-7	F-6	F-5	F-4	F-3	F-2	Brote semana (1)	Semana (0)	Datos de salida
					Semanas del brote				S
			Semanas antes del brote						N
		Dos semanas antes del brote							N

C. Datos de entrada (días de la semana)									
F-8	F-7	F-6	F-5	F-4	F-3	Brote semana (2)	Semana (1)	Semana (0)	Datos de salida
			Semanas del brote						S
		Semanas antes del brote							N
	Dos semanas antes del brote								N

D. Datos de entrada (días de la semana)									
F-8	F-7	F-6	F-5	F-4	Brote semana (3)	Semana (2)	Semana (1)	Semana (0)	Datos de salida
		Semanas del brote							S
	Semanas antes del brote								N
Dos semanas antes del brote									N

Figura 1. Representación visual de la ubicación de los datos de entrada en el clasificador bayesiano ingenuo en relación con las semanas consideradas. **A.** Semana 0. **B.** Semana 1. **C.** Semana 2. **D.** Semana 3. S: sí ocurrió el brote. N: no ocurrió el brote

años x 7 días x 4 semanas) para la clase S, y 252 (6 años x 7 días x 6 semanas) para la clase N por cada semana de antelación. En los modelos en los cuales se utilizaron los datos de cuatro años, fueron 112 valores (4 años x 7 días x 4 semanas) para la clase S y 168 (4 años x 7 días x 4 semanas) para la clase N por cada semana de antelación; y para el modelo en el que se utilizaron los datos de dos años fueron 56 valores (2 años x 7 días x 4 semanas) para la clase S y 84 (2 años x 7 días x 4 semanas) para la clase N por cada semana de antelación.

Se indujo un único clasificador bayesiano ingenuo para cada posible combinación (255 modelos) de las variables climatológicas en cada temporada de dos, cuatro y seis años, y para los modelos de 0, 1, 2 y 3 semanas antes del inicio del brote, se generaron 3.060 clasificadores. Los resultados de los rendimientos de cada uno de los 3.060 modelos incluyeron los valores de la sensibilidad y la especificidad, para los periodos de dos, cuatro y seis años.

Determinación de la semana del brote

Era necesario, en primer lugar, determinar un estándar de referencia para la fecha de inicio del brote, con el fin de analizar los modelos de cada año, para lo cual se necesitaban las gráficas de los casos de infección por semanas epidemiológicas en cada año.

Los datos de la infección por RSV en Bogotá se tomaron de los registros de vigilancia centinela de la circulación de virus respiratorios a cargo de las instituciones seleccionadas de común acuerdo por la Secretaría Distrital de Salud y el Instituto Nacional de Salud de Colombia para la vigilancia de las enfermedades respiratorias. En el Laboratorio de Salud Pública de la Secretaría Distrital de Bogotá, se diagnostican las infecciones por los principales virus respiratorios (influenza A, influenza B, sincitial respiratorio, adenovirus y parainfluenza 1, 2 y 3) mediante la técnica de inmunofluorescencia indirecta. Con base en dichos resultados se determinan, principalmente, el número de casos

y el porcentaje de resultados positivos del RSV por semana epidemiológica en niños menores de cinco años, así como el porcentaje de RSV entre los virus respiratorios detectados por semana epidemiológica en la vigilancia centinela.

En la figura 2 se presentan los casos de infección por RSV en niños menores de cinco años de edad en Bogotá, datos que sirvieron para estimar las siguientes semanas de inicio del brote en cada año: semana 13 de 2005, semana 12 de 2006, semana 14 de 2007, semana 13 de 2008, semana 13 de 2009 y semana 12 de 2010.

Es importante mencionar que los datos históricos que se presentan en la figura 2 demostraron una cierta coherencia en cuanto a la aparición del brote. Sin embargo, en el futuro podría no ser así debido a los cambios climáticos, como ha sucedido con la aparición de enfermedades en nuevas regiones del mundo.

A manera de resumen de la metodología utilizada en este trabajo, se elaboró un diagrama de proceso o de flujo para facilitar la comprensión del método (figura 3).

Resultados

Clasificación de los modelos

Para decidir cuáles modelos eran mejores para predecir el brote, se usó la gráfica de la fracción de falsos positivos (complemento de la especificidad) contra la de la fracción de verdaderos positivos (sensibilidad). En la figura 4 se presentan los tipos de resultados o puntos, los cuales representan todos los valores obtenidos en relación con las fracciones de verdadero positivo y falso positivo.

En la curva ROC (17) se evidenció que la zona de aceptación correspondía a la región sobre la diagonal y, la de rechazo, a la ubicada por debajo de la diagonal. En el cuadro 2 se identifican los modelos aceptados y rechazados, y se observan las veces que aparece cada variable climática en los modelos.

En la última columna del cuadro 2 se presentan los modelos aceptados de la semana cero de cada estudio (6, 4 y 2 años), los cuales resultaron ser los más aceptados en comparación con los de otras semanas. Es importante mencionar que, entre los modelos que se ubicaron en la zona de aceptación

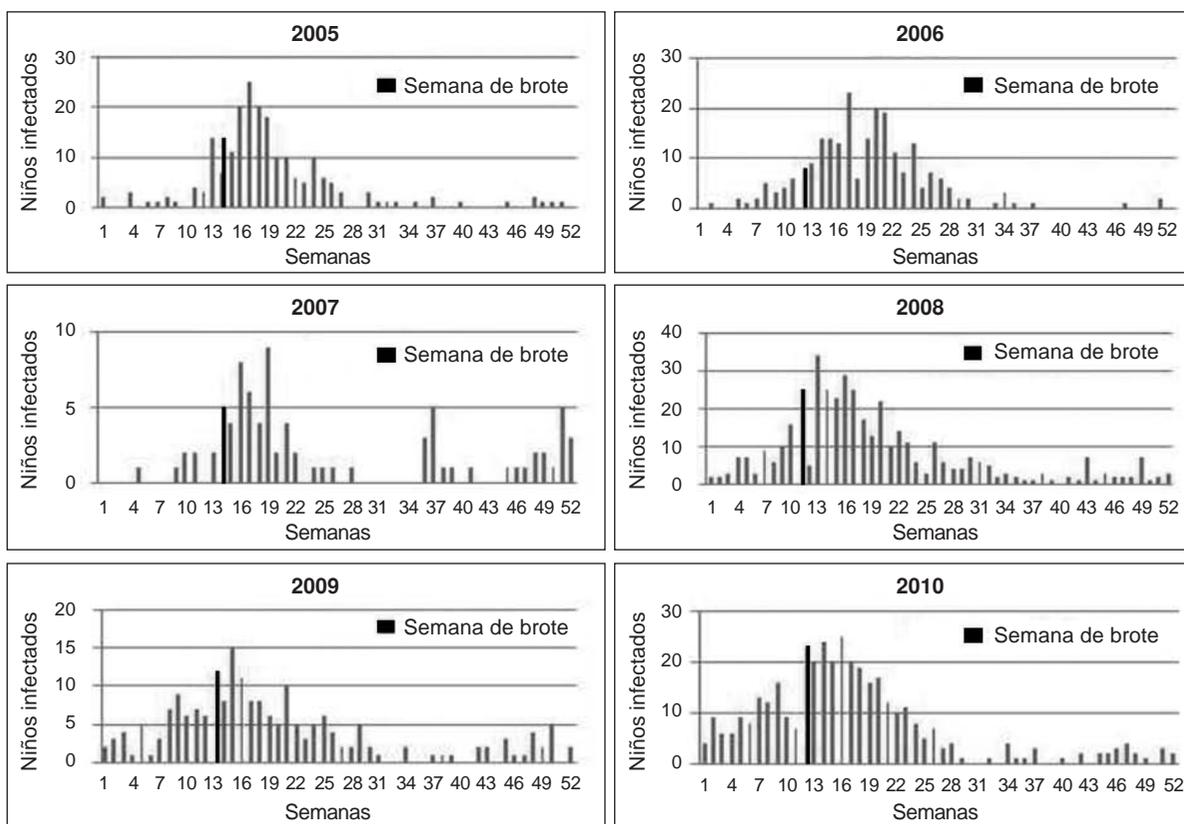


Figura 2. Gráficas utilizadas para analizar el comienzo de la semana del brote de la infección por el RSV

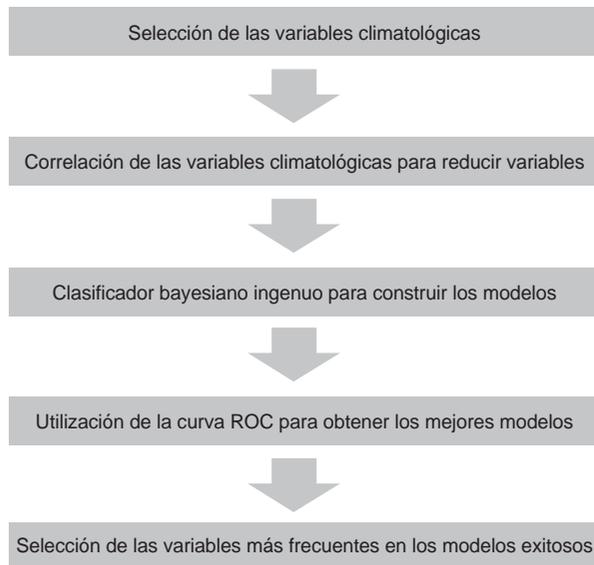


Figura 3. Diagrama del proceso de aplicación de la metodología utilizada

(sobre la diagonal de la curva ROC), para el análisis de resultados solo se utilizaron aquellos ubicados en el noreste de la gráfica de la fracción de falsos positivos contra la fracción de verdaderos positivos, es decir, el cuadro superior izquierdo de cada gráfica (9). De esta forma, se seleccionaron los modelos de mayor rendimiento para predecir el brote de la infección por RSV. Se asumió que los modelos de mejor rendimiento estaban en el área del cuadrado con los vértices (0,5, 0,5), (0,5,1), (0,0,5) y (0,1) (zona noreste).

Rendimiento individual de las variables climatológicas

La importancia relativa de una variable meteorológica en los modelos de mejor desempeño, se definió como el porcentaje de aparición de la variable en dichos modelos (zona noreste). En el cuadro 3 se presentan los porcentajes para cada variable climatológica. En este trabajo se consideró que un modelo tenía buen desempeño si los resultados estaban por encima de 50 % de sensibilidad y especificidad.

Para los modelos que consideraban el periodo de 2005 a 2010, el porcentaje mínimo de sensibilidad fue de 50 % y, el de especificidad, de 53 %. De esta forma, la humedad mínima fue la variable que apareció con mayor frecuencia en los modelos de buen rendimiento en la predicción en la misma semana del brote (semana 0), con una semana de antelación (semana 1) y en las dos semanas

anteriores al inicio del brote (semana 2). Sin embargo, fue menos representativa en los modelos de mejor rendimiento para predecir el brote con tres semanas de antelación.

Por otra parte, la temperatura mínima y el viento fueron las variables con las que se obtuvieron los resultados más precisos en los modelos de mejor rendimiento (zona noreste), los cuales predijeron el brote con tres semanas de antelación.

Por último, se observó que las variables con un promedio mayor de 50 % en todos los mejores modelos de las semanas fueron la temperatura máxima, la temperatura mínima, el viento y la humedad mínima, y que la variable con el menor promedio en todos los modelos de buen rendimiento fue la precipitación.

Análogamente, se estableció un modelo de buen desempeño utilizando los datos de cuatro años, así como aquellos con valores de sensibilidad superiores a 54 % y una especificidad mayor de 53 %. Para los modelos de mejor desempeño, la presión atmosférica y la humedad mínima fueron las variables más frecuentes para predecir la semana del brote. La humedad mínima fue la variable más frecuente en los modelos de mejor rendimiento, los cuales predijeron el inicio del brote con una y dos semanas de antelación. En los modelos que predijeron el inicio de la epidemia con tres semanas de antelación, las variables más frecuentes fueron la temperatura promedio y el viento. Por último, se observó que las variables con un promedio superior a 50 % en todos los modelos de semana fueron la temperatura promedio, la temperatura máxima, la temperatura mínima, el viento y la humedad mínima. Por otra parte, la variable con el menor promedio en todos los modelos fue la precipitación.

Para determinar los mejores modelos basados en el periodo de dos años, se asumió que debían superar valores de 50 % de sensibilidad y de 58 % de especificidad. Para la semana de inicio del brote (semana 0), de los 255 modelos, 250 podían considerarse como de buen rendimiento en la predicción; por lo tanto, no se pudo determinar la variable climatológica de mayor relación con el brote de infección por RSV. La temperatura mínima fue la variable mejor representada (69 %) en los mejores modelos para la predicción una semana antes del brote. En el modelo de predicción con dos semanas de antelación al inicio del brote, la variable más frecuente fue la velocidad del viento, con 79 %. La presión atmosférica, con

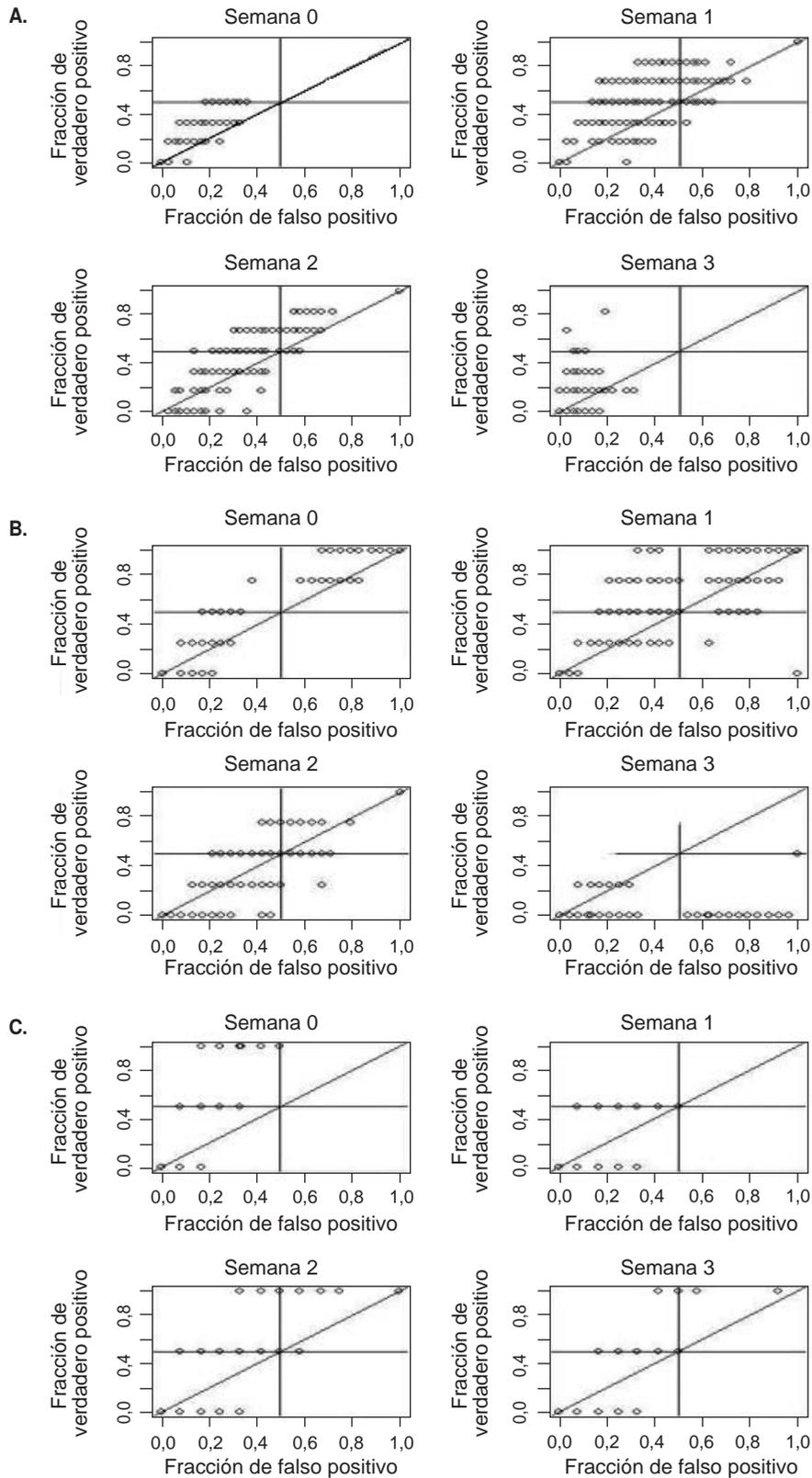


Figura 4. Gráfica de la fracción de verdaderos positivos (eje Y) contra la fracción de falsos positivos (eje X). **A.** Se presentan las veces que se repite cada punto en el periodo de seis años. **B.** Se presentan las veces que se repite cada punto en el periodo de cuatro años. **C.** Se presentan las veces que se repite cada punto en el periodo de dos años.

80 %, fue la más frecuente en los modelos de acierto en la tercera semana antes del comienzo de la epidemia. Por último, se observó que las variables con un promedio mayor de 50 % en todas las semanas fueron la temperatura máxima, la temperatura mínima, la velocidad del viento, la humedad mínima y la presión atmosférica. Por otra parte, la variable con menor promedio en todas las semanas fue la precipitación.

Al considerar el promedio para las variables climatológicas en todos los modelos de todas las temporadas y semanas, se encontró que la variable de humedad mínima aparecía en 62 % de los modelos de mejor rendimiento, lo cual demostró una correlación con el comienzo del brote. La variable de velocidad del viento apareció en 59 % de los modelos de mejor rendimiento. Por último, la variable de temperatura mínima tuvo un porcentaje de aparición de 57 % en los modelos de mejor desempeño. Es importante resaltar que la

precipitación apareció solo en 40 % de los modelos con buen rendimiento de predicción, lo cual indicó que no tuvo tanta correlación con la aparición de la infección por RSV como las variables climatológicas mencionadas antes.

Discusión

En este estudio se construyeron varios modelos basados en los clasificadores bayesianos ingenuos para predecir la semana de inicio del brote de la infección por RSV en Bogotá, para lo cual se utilizaron los datos climatológicos y el número de casos en niños menores de cinco años a partir de 2005 y hasta 2010. Se utilizaron tres periodos para evaluar los distintos modelos: dos años (2009-2010), cuatro años (2005-2008) y seis años (2005-2010).

Los mejores modelos para predecir el inicio del brote de RSV se obtuvieron para los periodos de seis (2005-2010) y dos (2009-2010) años. Es

Cuadro 2. Modelos aceptados y rechazados de los clasificadores de la semana 0 utilizando el periodo de seis, cuatro y dos años

Modelos	Temperatura promedio	Temperatura máxima	Temperatura mínima	Humedad máxima	Humedad mínima	Velocidad del viento	Presión atmosférica	Precipitación	Total
Seis años (2005-2010)									
Aceptados	115	116	108	112	128	120	123	118	226
Rechazados	13	12	20	16	0	8	5	10	29
Cuatro años (2005-2008)									
Aceptados	115	106	116	117	115	117	127	113	219
Rechazados	12	22	12	11	13	11	1	15	36
Dos años (2009-2010)									
Aceptados	127	128	126	128	128	123	126	128	250
Rechazados	1	0	2	0	0	5	2	0	5

Cuadro 3. Porcentaje de los modelos de las temporadas de seis y cuatro años que incluían cada variable en los modelos de mejor rendimiento (sensibilidad mínima de 50 % y especificidad mínima de 53 %)

Semana predictora	Temperatura promedio	Temperatura máxima	Temperatura mínima	Humedad máxima	Humedad mínima	Velocidad máxima	Presión Atmosférica	Precipitación
Seis años (sensibilidad mínima de 50 % y especificidad mínima de 53 %)								
Semana 0	59	51	55	54	87	49	57	51
Semana 1	46	56	45	42	78	42	45	55
Semana 2	43	54	53	53	76	66	51	36
Semana 3	0	50	100	0	33	83	17	0
Cuatro años (sensibilidad mínima de 50 % y especificidad mínima de 54 %)								
Semana 0	61	49	54	7	68	46	70	54
Semana 1	46	56	51	6	72	46	52	45
Semana 2	44	64	56	48	86	50	52	49
Semana 3	100	53	40	13	47	93	7	47
Cuatro años (sensibilidad mínima de 50 % y especificidad mínima de 58 %)								
Semana 0	51	51	50	51	51	49	50	51
Semana 1	43	57	69	21	59	41	39	40
Semana 2	48	51	50	48	44	79	56	40
Semana 3	46	44	58	67	49	58	80	12

importante resaltar que para determinar el mejor modelo se aplicó la metodología de curvas ROC, la cual permite determinar la idoneidad de un modelo con base en el porcentaje de falsos positivos.

Para la determinación de las variables climatológicas más relevantes en cuanto al inicio del brote de RSV, se tomó en cuenta la frecuencia (porcentaje) de aparición de cada variable en los mejores modelos según el método de las curvas ROC. De esta forma, se obtuvo que la humedad mínima resultaba ser la variable más significativa, al aparecer con un porcentaje de 62 % en los modelos de mejor rendimiento. Las variables climatológicas más importantes resultaron ser, en su orden, la velocidad del viento y la temperatura mínima, con porcentajes de presencia en los mejores modelos de 59 y 57 %, respectivamente. Debe mencionarse que estos resultados coinciden con los de otros trabajos (17,20), en los cuales se encontró que la humedad absoluta y la temperatura eran variables climatológicas correlacionadas con la prevalencia de RSV en Ciudad de México, Nueva Delhi, Buffalo, Miami, Houston, Tucson, Santiago de Chile, Winnipeg y Bethel (Alaska).

En trabajos anteriores se ha propuesto que la transmisión del RSV es posible mediante gotas grandes de aerosol o fómites (21,22). Además, se ha demostrado que las grandes partículas de aerosoles tienen una estabilidad máxima a una humedad de 90 % y una mínima con menos de 20 % de humedad (23). En su conjunto, estos resultados permiten explicar la relación de la humedad con el inicio del brote de RSV en Bogotá. Por otra parte, es importante mencionar que en Ciudad de México se encontró que el punto de rocío se correlacionaba con la actividad del RSV (20), lo cual es de resaltar porque Ciudad de México presenta algunas similitudes con Bogotá, y el punto de rocío está muy relacionado con la humedad. Los resultados de este trabajo coinciden parcialmente con los del estudio en Ciudad de México, donde no se encontró correlación con la temperatura (20).

Por otra parte, los resultados de este estudio en cuanto a la importancia de la humedad y la temperatura mínimas concuerdan con el trabajo de Welliver (22), en el cual se encontró que la humedad se correlacionaba con la aparición de brotes del RSV y que, particularmente en lugares fríos, se correlacionaba inversamente con la temperatura, lo cual puede ser el caso de Bogotá.

Es importante mencionar que la precipitación no estuvo entre las variables significativas para la predicción del inicio del brote, lo cual coincide con los resultados de otros estudios del RSV en diversas regiones (20,22,23). Al comparar los resultados de este estudio con los del realizado en el condado de Salt Lake City en Utah (10), puede observarse que la temperatura y la velocidad del viento fueron significativas en ambos estudios, lo cual le da más coherencia y validez a las presentes conclusiones. En ese trabajo se encontró que estas dos eran las mejores variables predictoras. En este trabajo, la humedad mínima fue la variable con mayor porcentaje de aparición en los mejores modelos, lo cual representa una diferencia con el trabajo en el condado de Salt Lake City en Utah. Sin embargo, la humedad mínima y la máxima no se descartaron porque aparecieron en varios de los mejores modelos, pero no con la frecuencia de la velocidad del viento y la temperatura. Es importante resaltar que estas variables climatológicas se encuentran relacionadas en distintos grados, por lo que puede existir correlación en los resultados de ambos estudios. En un trabajo en los Países Bajos (23), se encontró que la temperatura mínima y la humedad relativa se relacionaban con la actividad del RSV, lo cual concuerda en gran parte con nuestros resultados.

En un reciente estudio se sugiere que las temperaturas bajas pueden activar los viriones inactivos y, de esta forma, dar inicio al brote epidémico de RSV en la población (24). Este resultado coincide con el resultado de este trabajo en relación con la variable de temperatura mínima. De igual forma, en un estudio transversal reciente con datos de Bogotá, se encontró que la temperatura promedio y la lluvia estaban independientemente asociadas con el número de niños hospitalizados con infección por el RSV (25). Por otra parte, en un estudio con datos de la ciudad de Fukuoka en Japón, se reveló evidencia cuantitativa del número de casos entre el 2006 y el 2012, el cual aumentó significativamente con el rango de la temperatura diaria (26).

En cuanto a los mejores modelos de las semanas de anticipación para la predicción del brote en este estudio, se encontró que aquellos basados en los datos climatológicos correspondientes a las cuatro semanas previas al inicio del brote eran los más efectivos en la predicción. En el estudio realizado en el condado de Salt Lake City en Utah (10), se pudo observar que los modelos que utilizaron las variables climatológicas presentes

en las tres semanas previas al brote resultaron los mejores modelos predictivos. Teniendo en cuenta la diferencia de solo una semana, los resultados apuntan a no tener en cuenta los datos climatológicos de la semana anterior al brote.

Por otra parte, en el estudio de la actividad del RSV en los Países Bajos, se encontró que la temperatura mínima y la humedad relativa en las semanas previas se correlacionaban con la aparición del brote de RSV (23). Estos resultados parecen correlacionarse con el hecho de que el RSV tiene un periodo de incubación de una semana aproximadamente, y la aparición del brote epidémico toma cierto tiempo después de que se infectan las primeras personas. De esta forma, se puede sugerir a las autoridades sanitarias la vigilancia de las variables climatológicas tan pronto como sea posible, con el fin de lograr una mayor probabilidad de predicción.

En este trabajo se utilizaron los clasificadores bayesianos ingenuos para determinar los mejores modelos para predecir la semana de inicio del brote de infección por RSV. Los resultados sugieren que los modelos que utilizaron la humedad mínima, la velocidad del viento y la temperatura mínima fueron los de mayor potencial para la predicción del inicio del brote de RSV anualmente. Es importante señalar que estas variables climatológicas están correlacionadas de alguna forma. Se requieren más estudios que correlacionen los valores de estas variables climatológicas con el inicio del brote del RSV para llegar a conclusiones específicas de predicción. Por lo tanto, se recomienda vigilar las variables climatológicas relevantes en Bogotá, con el fin de predecir el inicio del brote de RSV en un futuro y tomar las medidas pertinentes en salud pública y prevención.

Es importante reconocer que la relación entre las variables climatológicas y la aparición del brote del RSV puede verse influenciada por factores confusos. Dos aspectos interesantes son que, la mayoría de las veces, la transmisión del RSV ocurre en recintos cerrados y que los patrones sociales están influidos por el clima. Por lo tanto, la asociación entre el clima y la aparición de brotes de RSV podría ser secundaria en algunos casos (20).

Por último, en futuros estudios se sugiere incluir las bases de datos de más años, así como otras variables relacionadas con la contaminación del aire (concentración de CO y SO₂) para, así, mejorar los modelos de predicción. De igual forma, se deberá incluir la variable de exposición a la luz

ultravioleta, ya que se ha sugerido que esta exposición puede inactivar el virus en el ambiente o afectar la sensibilidad frente a él mediante la alteración de la resistencia del huésped por incremento en los niveles de vitamina D (20).

Conflicto de intereses

Los autores manifiestan no tener conflicto de intereses que ocasionen algún sesgo en el manejo y la interpretación de la información proporcionada en este estudio.

Financiación

Este trabajo fue financiado parcialmente por el Consejo de Desarrollo Científico, Humanístico, Tecnológico y de las Artes (CDCHTA) de la Universidad de Los Andes, mediante los proyectos I-1331-12-05-EM y I-1289-11-05-A.

Referencias

1. **Benguigui Y, López FJ, Sdhmunis G, Yunes J.** Infecciones respiratorias en niños. Serie HCT/AIEPI-1. Washington, D.C.: Organización Panamericana de la Salud; 1997. p. 496.
2. **López-Cuadrado T, de Mateo S, Jiménez-Jorge S, Savulescu C, Larrauri A.** Influenza-related mortality in Spain, 1999-2005. *Gac Sanit.* 2012;26:325-9. <http://dx.doi.org/10.1016/j.gaceta.2011.09.033>
3. **van Woensel JB, van Aaldersden WM, Kimpfen JL.** Viral lower respiratory tract infection in infants and young children. *BMJ.* 2003;327:36-40. <http://dx.doi.org/10.1136/bmj.327.7405.36>
4. **Aranda-Lozano DF, González-Parra GC, Querales J.** Modelamiento de la transmisión del virus respiratorio sincitial (VRS) en niños menores de cinco años. *Rev Salud Pública.* 2013;15:637-47.
5. **Rodríguez J, Prieto S, Correa C, Arnold Y, Álvarez L, Bernal P, et al.** Dynamics of the dengue epidemic in Colombia: Predictions of the epidemic trajectory. *Rev Fac Med.* 2013;21:38-45.
6. **Núñez LM, Aranda DF, Jaramillo AC, Moyano LF, Osorio ED.** Chronology of a pandemic: The new influenza A (H1N1) in Bogotá, 2009-2010. *Rev Salud Pública.* 2011;13:480-91. <http://dx.doi.org/10.1590/S0124-00642011000300010>
7. **González-Parra GC, Villanueva-Micó RJ, Segovia L.** Dinámica del virus pandémico AH1N1/09 en la población de Venezuela. *Interciencia.* 2012;37:306-11.
8. **González-Parra G, Villanueva RJ, Ruiz-Baragaño J, Moraño JA.** Modelling influenza A (H1N1) 2009 epidemics using a random network in a distributed computing environment. *Acta Trop.* 2015;143:29-35. <http://dx.doi.org/10.1016/j.actatropica.2014.12.008>
9. **Adarme LV, Castellanos JE.** Diagnóstico virológico de la infección por virus sincitial respiratorio. *Revista Salud Bosque.* 2015;3:23-36.
10. **Walton NA, Poynton MR, Gesteland PH, Maloney C, Staes C, Facelli JC.** Predicting the start week of respiratory

- syncytial virus outbreaks using real time weather variables. *BMC Med Inform Decis Mak.* 2010;10:68. <http://dx.doi.org/10.1186/1472-6947-10-68>
11. **Russell S, Norvig P.** *Intelligence A. A modern approach.* Artificial Intelligence. Prentice-Hall, Englewood Cliffs. 1995; 25:27.
 12. **Domingos P, Pazzani M.** On the optimality of the simple Bayesian classifier under zero-one loss. *Mach Learn.* 1997; 29:103-30. <http://dx.doi.org/10.1023/A:1007413511361>
 13. **Noguera T.** Metodología ROC en la evaluación de medidas antropométricas como marcadores de la hipertensión arterial. Aplicación a población gallega adulta (tesis). Santiago de Compostela: Universidad Santiago de Compostela; 2010.
 14. **Lim TS, Loh WY, Shih YS.** An empirical comparison of decision trees and other classification methods. Technical report 979. Madison, WI: University of Wisconsin; 1997.
 15. **Noyola DE, Mandeville PB.** Effect of climatological factors on respiratory syncytial virus epidemics. *Epidemiol Infect.* 2008;136:1328-32. <http://dx.doi.org/10.1017/S0950268807000143>
 16. **Omer SB, Sutanto A, Sarwo H, Linehan M, Djelantik IG, Mercer D, et al.** Climatic, temporal, and geographic characteristics of respiratory syncytial virus disease in a tropical island population. *Epidemiol Infect.* 2008;136:1319-27. <http://dx.doi.org/10.1017/S0950268807000015>
 17. **Yusuf S, Piedimonte G, Auais A, Demmler G, Krishnan S, van Caesele P, et al.** The relationship of meteorological conditions to the epidemic activity of respiratory syncytial virus. *Epidemiol Infect.* 2007;135:1077-90. <http://dx.doi.org/10.1017/S095026880600776X>
 18. **Fawcett T.** An introduction to ROC analysis. *Pattern Recognit Lett.* 2006;27:861-74. <http://dx.doi.org/10.1016/j.patrec.2005.10.010>
 19. **Ting SL, Ip WH, Tsang AH.** Is naive Bayes a good classifier for document classification? *International Journal of Software Engineering and its Applications.* 2011;5:37-46.
 20. **Welliver R.** The relationship of meteorological conditions to the epidemic activity of respiratory syncytial virus. *Paediatr Respir Rev.* 2009;10(Suppl.1):6-8. [http://dx.doi.org/10.1016/S1526-0542\(09\)70004-1](http://dx.doi.org/10.1016/S1526-0542(09)70004-1)
 21. **Tang JW, Loh TP.** Correlations between climate factors and incidence –a contributor to RSV seasonality. *Rev Med Virol.* 2014;24:15-34. <http://dx.doi.org/10.1002/rmv.1771>
 22. **Welliver RC.** Temperature, humidity, and ultraviolet B radiation predict community respiratory syncytial virus activity. *Pediatr Infect Dis J.* 2007;26(Suppl.11):S29-35. <http://dx.doi.org/10.1097/INF.0b013e318157da59>
 23. **Meerhoff TJ, Paget JW, Kimpfen JL, Schellevis F.** Variation of respiratory syncytial virus and the relation with meteorological factors in different winter seasons. *Pediatr Infect Dis J.* 2009;28:860-6. <http://dx.doi.org/10.1097/INF.0b013e3181a3e949>
 24. **Stewart PD.** Seasonality and selective trends in viral acute respiratory tract infections. *Med Hypotheses.* 2016;86:104-19. <http://dx.doi.org/10.1016/j.mehy.2015.11.005>
 25. **Rodríguez-Martínez CE, Sossa-Briceño MP, Acuña-Cordero R.** Relationship between meteorological conditions and respiratory syncytial virus in a tropical country. *Epidemiol Infect.* 2015;143:2679-86. <http://dx.doi.org/10.1017/S0950268814003793>
 26. **Onozuka D.** The influence of diurnal temperature range on the incidence of respiratory syncytial virus in Japan. *Epidemiol Infect.* 2015;143:813-20. <http://dx.doi.org/10.1017/S0950268814001575>