MEDwave

Temas y controversias en bioestadística
Medwave 2015 Ene;15(1):e6075 doi: 10.5867/medwave.2015.01.6075

La regresión logística frente a una red bayesiana divergente

Logistic regression against a divergent Bayesian network.

Autor: Noel Antonio Sánchez Trujillo**[1]**

Filiación:
[1]Universidad de Antioquia, Medellín, Colombia

E-mail: asan@une.net.co

Citación: Sánchez NA. Logistic regression against a divergent Bayesian network. *Medwave.* 2015 Ene;15(1):e6075 doi: 10.5867/medwave.2015.01.6075
Fecha de envío: 7/1/2015
Fecha de aceptación: 16/1/2015
Fecha de publicación: 3/2/2015
Origen: no solicitado
Tipo de revisión: con revisión por dos pares revisores externos, a doble ciego

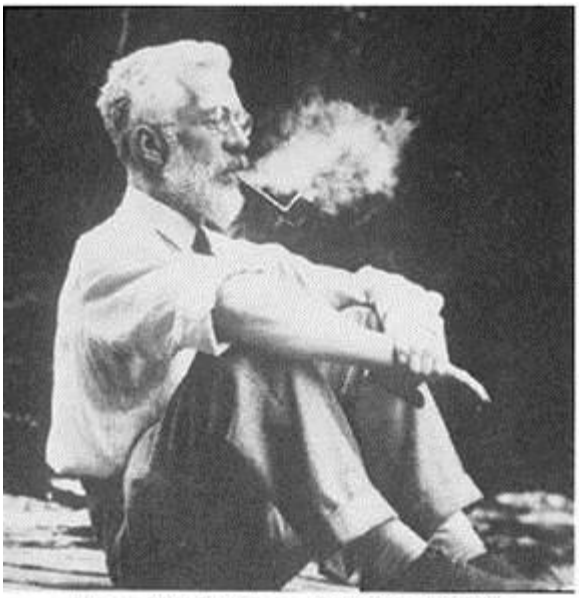
Resumen

Este artículo aborda dos herramientas estadísticas utilizadas en la predicción y la causalidad: la regresión logística y las redes bayesianas. A partir de datos tomados como ejemplo simulado de un estudio que evalúa factores que influyen en el padecimiento de enfisema pulmonar -donde se incluyen pigmentación en los dedos y tabaquismo- nos preguntamos: ¿es la pigmentación un factor de confusión, causal o predictivo?, ¿existe otro factor que, como el tabaquismo, confunde? O tal vez, ¿existe sinergia entre la pigmentación y el tabaquismo? En el caso de la predicción los resultados son similares con las dos técnicas, en cambio en la causalidad tienen diferencias. Se concluye que en la toma de decisiones es mejor la suma de una herramienta estadística usada con sentido común y evidencias anteriores, que toman años e incluso siglos para consolidarse, que el empleo automático y exclusivo de recursos estadísticos.

Abstract

This article is a discussion about two statistical tools used for prediction and causality assessment: logistic regression and Bayesian networks. Using data of a simulated example from a study assessing factors that might predict pulmonary emphysema (where fingertip pigmentation and smoking are considered); we posed the following questions. Is pigmentation a confounding, causal or predictive factor? Is there perhaps another factor, like smoking, that confounds? Is there a synergy between pigmentation and smoking? The results, in terms of prediction, are similar with the two techniques; regarding causation, differences arise. We conclude that, in decision-making, the sum of both: a statistical tool, used with common sense, and previous evidence, taking years or even centuries to develop; is better than the automatic and exclusive use of statistical resources.

Introducción



Ronald Fisher (1890-1962)

Thomas Bayes (1702-1761)

En el segundo decenio del siglo pasado, apenas se sospechaba la relación entre fumar y cáncer pulmonar. Luego, en la década de los cincuenta, con los trabajos de Ernst Wynder (1922-1999), Evarts Graham (1883-1957) [1], Richard Doll (1912-2005) y Bradford Hill (1897-1991) [2], la sospecha cobró un interés que produjo agrios debates en la opinión pública. Se preguntaban si fumar producía cáncer. No, respondieron las tabacaleras y prestigiosos estadísticos como Joseph Berkson (1899-1982), Jersy Neyman (1894-1981) y R. A. Fisher. Este último, el máximo paladín, defendió su posición en un artículo de 1958 titulado *Cigarettes, cancer and statistics* [3], al que se sumaron más tarde dos artículos en la prestigiosa revista *Nature*titulados *Lung cancer and cigarettes* [3] y *Cancer and smoking* [3]. Pero la acumulación de evidencias a favor de la hipótesis de que el tabaquismo era un factor de riesgo para contraer cáncer pulmonar, hizo cambiar de opinión a Berkson y a Neyman; no así a Fisher, quien posó complacido e impasible, fumando su pipa y arrojando una humareda.

En 1959, revisando los trabajos públicos y las objeciones de Fisher, Berkson y Neyman, y del propio Tobacco Institute, el gran estadístico Jerome Cornfield (1912-1979), con cinco expertos del *National Cancer Institute*, de la *American Cancer Society* y del *Sloan-Kettering Institute*, publicó una nueva evidencia respaldando la hipótesis de que el hábito de fumar es una causa importante del aumento en la incidencia de cáncer de pulmón [4]. Ahora, al cabo de los años, se sabe que el hábito tabáquico no solo es un factor de riesgo del cáncer pulmonar, sino también de otros tipos de cáncer y del enfisema pulmonar. Que el tabaquismo pudiera ser causa de tantas enfermedades, le parecía increíble a Berkson y esto, en apariencia, le hizo dudar de la causalidad entre el cigarrillo y el problema pulmonar [5]. Hoy son utilizadas diversas técnicas estadísticas para la predicción y la causalidad, dos de las cuales queremos resaltar y comparar: la regresión logística y las redes bayesianas. La regresión logística es un tipo de análisis utilizado comúnmente para predecir el resultado de una variable categórica en función de variables independientes que pueden ser cualitativas o cuantitativas. Debida a Cornfield, Gordon y Smith [6], surge en la década de los 60 con la aparición de un trabajo sobre el riesgo de padecer una enfermedad coronaria, y su uso se fue incrementando desde la década de los 80 gracias a las facilidades computacionales disponibles desde entonces pero, sobre todo, gracias a David R. Cox (1924) y a su libro, de 1970, *The Analysis of Binary Data*. A propósito, Cox comentó hace un lustro, en la revista *International Journal of Epidemiology*, que el artículo de Jerome Cornfield [4] era de importancia fundamental para todos los interesados en la historia del que puede ser el problema más importante de la epidemiologia sobre las enfermedades no infecciosas [7].

Las redes bayesianas se remontan hacia la década de 1980. Permiten modelar un problema estadístico cualitativa y cuantitativamente, de forma que la parte cualitativa corresponde a un grafo dirigido acíclico conexo y la cuantitativa a una distribución de probabilidad condicionada para cada variable. El grafo lo conforman un conjunto de nodos (variables aleatorias), cuyos enlaces dirigidos se representan mediante flechas. El primer nodo es "padre" del segundo, que es su "hijo". Entre dos nodos consecutivos puede existir un enlace (flecha), que se llama camino. Cuando esto ocurre, tenemos un grafo anexo. Si recorremos el camino siguiendo la dirección de los enlaces y podemos volver al punto de partida, tenemos un ciclo. Los grafos dirigidos acíclicos son aquellos que no tienen

ciclos. Las dos redes bayesianas médicas más famosas [8], aunque no hayan resuelto problemas reales, han sido útiles con fines ilustrativos. Estas son: la primera red bayesiana médica construida por G. F. Cooper en la Universidad de Stanford [9], conformada por cinco nodos referidos al tumor cerebral; y la red bayesiana del diagnóstico diferencial entre tuberculosis, bronquitis y cáncer pulmonar de Lauritzen y Spiegelhalter conformada, ya no por cinco sino por ocho nodos [10].

Una ilustración

Para ilustrar, con un caso simulado, las semejanzas y las diferencias entre una regresión logística y una red bayesiana, nos sirven los datos proporcionados por Silva [11], adaptados, aumentados y resumidos en la tabla 1.

Tabaquismo (hábito de fumar)						Total
		-t		+t		
		Enfisema		Enfisema		
		+e	-e	+e	-e	
Pigmentación en los dedos	+p	18	162	2250	17905	20335
	-p	54	954	18	252	1278
Total		72	1116	2268	18157	21613

Los signos + y – representan presencia y ausencia respectivamente.

Las letras t, e y p aluden a tabaquismo, enfisema pulmonar y pigmentación en los dedos respectivamente.

Tabla 1. Resultados de la asociación existente entre la pigmentación de los dedos como factor de riesgo y el enfisema pulmonar, por tabaquismo.

El problema

Al encontrar una asociación entre la pigmentación amarilla de los dedos y el enfisema pulmonar (P→E), con una razón de oportunidades (*odds ratio*, OR) de 2,10 y con un intervalo de confianza de 95% entre 1,65 y 2,68, cabe preguntarse: ¿es la pigmentación de los dedos un factor de confusión, causal o predictivo?, ¿existe otro factor que, como el tabaquismo, confunde? O bien, ¿tal vez se produce una sinergia entre la pigmentación y el tabaquismo?

Confusión

La relación existente entre la exposición que se estudia (X) y la enfermedad que resulta (Y) puede estar afectada por la superposición de un factor (Z), extraño, no tenido en cuenta, denominado factor de confusión. Una variable puede considerarse factor de confusión cuando cumple, tres premisas:

1. El factor de confusión (Z) nunca es un paso intermedio entre la exposición (X) y la enfermedad (Y).
2. El factor de confusión (Z) es un factor de riesgo para la enfermedad (Y), incluso en las personas que no tienen el factor exposición (-X).
3. El factor de confusión (Z) es un factor asociado a la exposición (X), no necesariamente factor de riesgo para esta, incluso en las personas que no tienen la enfermedad (-Y).

Luego, ¿la pigmentación de los dedos (P) es un factor de confusión para evaluar la asociación entre el tabaquismo (T) y el enfisema pulmonar (E)? A primera vista parecería que la respuesta es sí, puesto que tener los dedos pigmentados está asociado tanto con el tabaquismo como con el enfisema pulmonar; pero la pigmentación está asociada con el enfisema sólo de forma secundaria a su asociación con el tabaquismo, un factor de confusión debe ser predictivo de la ocurrencia de la enfermedad aparte de su asociación con la exposición, es decir, incluso entre los individuos no expuestos. Si la pigmentación fuese un factor predictivo de enfisema también entre los no fumadores, entonces sí sería un factor de confusión.

En cambio, si se quisiera evaluar la relación causal entre la pigmentación de los dedos y el enfisema, el tabaquismo sí cumple con las tres premisas como factor de confusión. Debido a que, nunca se daría una relación P→T→E (el tabaquismo no es un paso intermedio) y, siempre T→E (el tabaquismo es un factor de riesgo, incluso en las personas no pigmentadas) y T→P aunque no haya enfisema (el tabaquismo está asociado con la pigmentación, incluso en los pacientes sin enfisema pulmonar).

Interacción

La interacción se presenta cuando el efecto de un factor X (pigmentación en los dedos), sobre el desenlace Y (enfisema pulmonar), depende de cuál sea el nivel del otro factor T (tabaquismo). Cuando la presencia de un factor aumenta el efecto de otro, decimos que la interacción es sinérgica. Introduciendo como variable dependiente al enfisema pulmonar y, como variables independientes, la pigmentación de los dedos y el tabaquismo, en el programa *EpiInfo*™ 7.1.4, la regresión logística presenta los resultados de la tabla 2.

Regresión Logística M -M				Regresión Logística M +M			
Variables	OR*	Coeficiente ß	P	Variables	OR*	Coeficiente ß	P
P	1,84	0,6116	0,001	P	1,96	0,6745	0,018
T	1,19	0,1715	0,366	T	1,26	0,2326	0,408
				P*T	0,90	-0,1095	0,771
Constante	*	-2,8569	0,000	Constante	*	-2,8717	0,000

**Odds Ratio* estimado

Tabla 2. Resultados de la regresión logística según un modelo no multiplicativo (M -M) y multiplicativo (M +M).

Podría pensarse en una posible interacción sinérgica entre el grado de pigmentación y el tabaquismo como factores de riesgo para el enfisema, relación dependiente del tipo de tabaco, con filtro o sin filtro, la manera y la frecuencia de inhalación y la duración del hábito. Aquí hay otro problema con respecto a la diferenciación entre la interacción aditiva y la multiplicativa, en el cual no ahondaremos por la brevedad del artículo. En todo caso, los resultados de la tabla 2 muestran que el coeficiente de interacción (p*t) es de -0,1095 (p=0,77). Puesto que no se puede rechazar la hipótesis de que el coeficiente es igual a cero, por tener un valor p mayor a 0,05 (valor subjetivo y dogmático), habitualmente se concluye que no existe interacción multiplicativa entre las variables; en este caso, entre tener los dedos pigmentados y el tabaquismo. Además, según la tabla 1, el estimado de la razón de oportunidades (*odds ratio*, OR_p), para sólo la pigmentación de los dedos, en ausencia de fumar, es:

$$\hat{OR}_p = \frac{18 \times 954}{162 \times 54} = 1{,}96$$

Por lo contrario, el OR_f, para sólo fumar, en ausencia de pigmentación, es:

$$\hat{OR}_f = \frac{18 \times 954}{252 \times 54} = 1{,}26$$

Y, el OR_{f*p}, combinando el fumar y la pigmentación de los dedos, es:

$$\hat{OR}_{f*p} = \frac{2250 \times 954}{17905 \times 54} = 2{,}22$$

Rothman proporciona una fórmula para calcular el exceso de riesgo (RERI) [12] debido a la interacción bajo un modelo aditivo:

$$RERI = 2{,}22 - 1{,}96 - 1{,}26 + 1 = 0$$

$$RERI = \hat{OR}_{f*p} - \hat{OR}_p - \hat{OR}_f + 1$$

Y, si realizamos un análisis estratificado, obtenemos un valor para Ji cuadrado (prueba de homogeneidad) de la diferencia de los *odds ratio* por estratos (interacción) de 0,928, concordante con la regresión logística. Luego, en este lugar y en un tiempo específico, no existe interacción ni aditiva ni multiplicativa entre el tabaquismo y la pigmentación.

Predicción

Cuando dos variables están relacionadas, es posible predecir una de ellas basándose en el conocimiento de la otra. Esta relación entre dos variables conlleva frecuentemente la implicación de que una variable es causa de la otra. De ahí que se suela pasar por alto el hecho de que las variables pueden no estar unidas causalmente, sino que varían juntas en virtud de un vínculo común con una tercera variable. Las probabilidades de tener enfisema las obtenemos, mediante la regresión logística, por medio de tres modelos, uno para cada una de las evidencias, basados en la ecuación de la función logística (calculadas en la tabla 3):

$$P(+e \mid p) = \frac{e^{-2{,}8184 + 0{,}7432 \times p}}{1 + e^{-2{,}8184 + 0{,}7432 \times p}}$$

1 y 2

$$P(+e \mid t) = \frac{e^{-2{,}7408 + 0{,}6607 \times t}}{1 + e^{-2{,}7408 + 0{,}6607 \times t}}$$

3 y 4

$$P(+e \mid p, t) = \frac{e^{-2{,}8569 + 0{,}6116 \times p + 0{,}1715 \times t}}{1 + e^{-2{,}8569 + 0{,}6116 \times p + 0{,}1715 \times t}}$$

5, 6, 7 y 8:

Entre los diversos modelos posibles, como puede observarse en la figura 1, se decide hacer una red bayesiana divergente, construida en el programa Elvira [13].

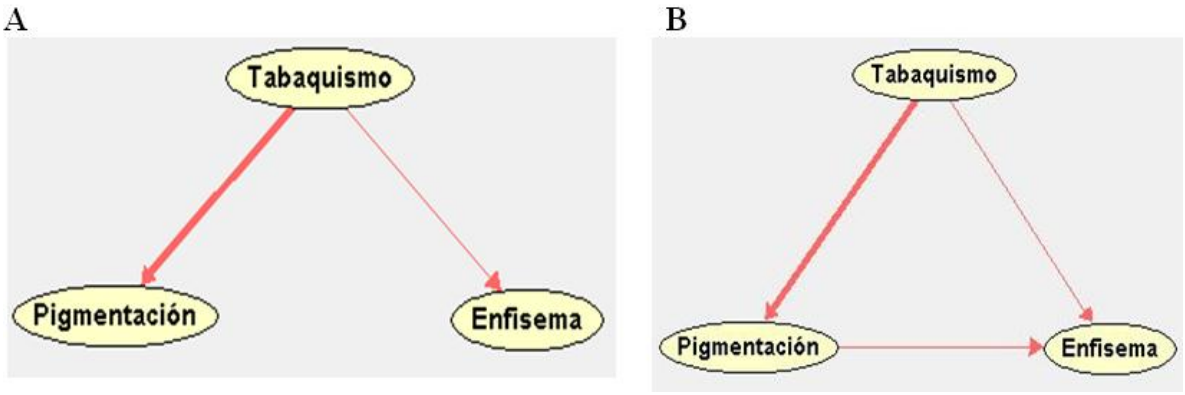


Figura 1. A red bayesiana divergente 3. B red bayesiana combinada 6 (Tabla 3).

En la figura 1, los nodos representan variables aleatorias y los arcos simbolizan relaciones de dependencia directa entre las variables. Con tres variables e, p y t, las redes bayesianas ofrecen tres tipos de conexiones básicas: seriales, T→P→E; convergentes, P→E← y, divergentes, P←T→E. Hay una cuarta conexión, la combinada, la cual presenta una relación directa entre todas las variables (Figura 1B). La figura 1A muestra una conexión divergente: tiene un nodo padre (tabaquismo) que proyecta sus arcos hacia dos hijos (la pigmentación y el enfisema). Es la conexión más apropiada para representar enfermedades o trastornos detectables por sus síntomas. Este programa, en modo de diseño y de inferencia, interpreta, mediante el grosor y el color del arco, el grado de asociación entre dos variables. Así, el mayor grosor indica que existe un mayor grado de asociación entre el tabaquismo y la pigmentación que entre el tabaquismo y el enfisema, o entre la pigmentación y el enfisema; una flecha roja indica que existe una relación positiva (azul indicaría relación negativa; púrpura, una relación intermedia; y negra, ausencia de relación).

Las probabilidades de tener enfisema para las redes bayesianas se obtienen mediante la fórmula de la probabilidad conjunta del modelo de la figura 1 y los datos de la tabla 1. Los resultados se resumen en la tabla 3:

Evidencias (1-8)	1. +p	2. -p	3. +t	4. -t	5. +t, +p	6. -t, -p	7. +t, -p	8. -t, +p
Técnica								
Regresión logística								
1. Sin interacción	0,112	0,056	0,111	0,061	0,112	0,054	0,064	0,096
2. Con interacción	0,112	0,056	0,111	0,061	0,112	0,054	0,067	0,100
Redes bayesianas								
Seriales*								
1.1 Serial 1	0,112	0,056	0,111	0,065	0,112	0,056	0,056	0,112
1.2 Serial 2	0,112	0,056	0,111	0,065	0,112	0,056	0,056	0,112
1.3 Serial 3	0,112	0,056	0,111	0,061	0,114	0,031	0,058	0,063
1.4 Serial 4	0,112	0,056	0,111	0,061	0,114	0,031	0,058	0,063
1.5 Serial 5	0,111	0,071	0,111	0,061	0,111	0,061	0,111	0,061
1.6 Serial 6	0,111	0,071	0,111	0,061	0,111	0,061	0,111	0,061
Divergentes †								
2.1 Divergente 1	0,112	0,056	0,111	0,065	0,112	0,056	0,056	0,112
2.2 Divergente 2	0,112	0,056	0,111	0,061	0,114	0,031	0,058	0,063
2.3 Divergente 3	0,111	0,071	0,111	0,061	0,111	0,061	0,111	0,061
Convergentes ‡								
3.1 Convergente 1	0,110	0,090	0,108	0,108	0,110	0,096	0,065	0,173
3.2 Convergente 2	0,111	0,066	0,109	0,097	0,112	0,054	0,066	0,100
3.3 Convergente 3	0,108	0,108	0,109	0,102	0,108	0,103	0,127	0,097
Combinadas §								
4.1 Combinada 1	0,112	0,056	0,111	0,061	0,112	0,054	0,067	0,100
4.2 Combinada 2	0,112	0,056	0,111	0,061	0,112	0,054	0,067	0,100
4.3 Combinada 3	0,112	0,056	0,111	0,061	0,112	0,054	0,067	0,100
4.4 Combinada 4	0,112	0,056	0,111	0,061	0,112	0,054	0,067	0,100
4.5 Combinada 5	0,112	0,056	0,111	0,061	0,112	0,054	0,067	0,100
4.6 Combinada 6	0,111	0,063	0,112	0,056	0,112	0,054	0,100	0,067

*serial 1: T→P→E, serial 2: E→P→T, serial 3: T→E→P, serial 4: P→E→T, serial 5: P→T→E, serial 6: E→T→P

† divergente 1: T→P←E; divergente 2: P←E→T; divergente 3: P←T→E

‡convergente 1: T→P←E; convergente 2: P→E←T; convergente 3: P→T←E

§ combinada 1: P(p,t,e)= P(p).P(e/p).P(t/p,e); combinada 2: P(p,t,e)=P(p).P(t/p).P(e/p,t); combinada 3: P(p,t,e)=P(e).P(p/e).P(t/p,e); combinada 4: P(p,t,e)=P(e).P(t/e).P(p/t,e); combinada 5: P(p,t,e)=P(t).P(e/t).P(p/t,e); combinada 6: P(p,t,e)=P(t).P(p/t).P(e/p,t)

Tabla 3. Probabilidad de tener enfisema dada(s) una(s) evidencia(s) según diferentes técnicas estadísticas.

Notemos que las diferencias de las probabilidades en las evidencias 1(+p) y 2(-p), para la regresión logística (P(+e/+p)=11,2% y P(+e/-p=5,6%) y en la red bayesiana divergente 3 (P(+e/+p)=11,1% y P(+e/-p =7,1%), no son amplias; sin embargo, la red bayesiana incluye una posibilidad no incluida en la regresión logística (Figura 2).

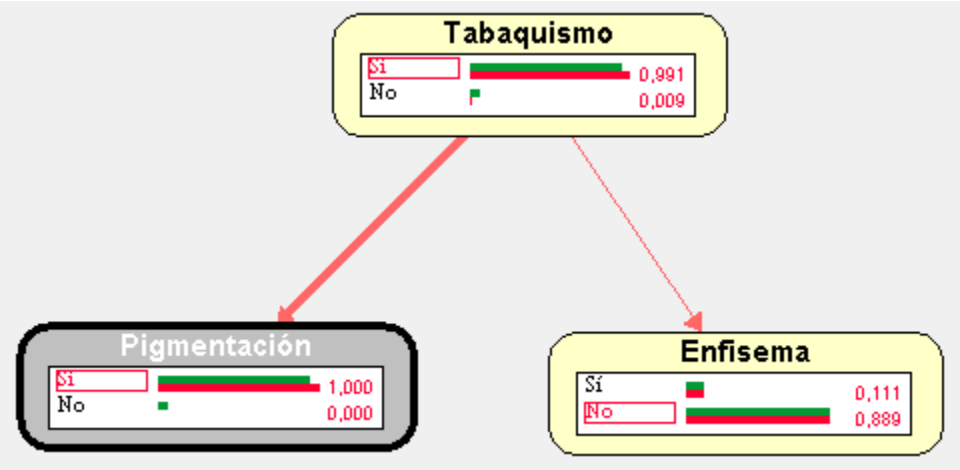


Figura 2. Red enfisema con evidencia de pigmentación.

En la figura 2 vemos como la red bayesiana permite hacer un recorrido entre la pigmentación y el enfisema, pasando por el tabaquismo. En otras palabras, introducir la información "un paciente tiene los dedos pigmentados", provoca un incremento en la suposición de que fuma, la cual se asocia al enfisema, y hace que la creencia en la presentación de esta enfermedad aumente. Pero podemos construir una red bayesiana adicional para la pigmentación y el enfisema (P→E) y, de esta forma, la predicción de las evidencias 1 y 2 con la regresión logística y la red bayesiana se iguala. Recordemos que los resultados de la tabla 3 en la regresión logística están basados en tres modelos y, en la red bayesiana, en uno. Los pronósticos con las evidencias 3 (+t) y 4 (-t) son idénticos en la regresión logística y la red bayesiana divergente 3 y con las evidencias 5 (+t, +p) y 6 (-t, -p) las diferencias de las probabilidades entre la regresión logística y la red bayesiana divergente 3 son mínimas. La diferencia la hacen las evidencias 7 y 8: P(+e/-t,+p) 9,6% y 6,1%; y P(+e/+t,-p) 6,4% y 11,1%, para la regresión logística y la red bayesiana divergente 3, respectivamente. Estos resultados muestran que puede haber una diferencia de hasta 58% (6,4/11,1) entre las dos herramientas.

Con respecto a la predicción, la regresión logística, multiplicativa y no multiplicativa, da valores similares porque no hay interacción entre el tabaquismo y la pigmentación con respecto a la aparición de enfisema (tabla 3). Con fines pragmáticos, si el investigador quiere construir una función que muestre cuán probable es que un individuo sea enfisematoso si tiene los dedos pigmentados y fuma, el modelo de regresión puede ser útil para calcular el riesgo. No obstante, hay diferencias pronósticas de hasta 58% en las evidencias 7 y 8 para la regresión logística y la red bayesiana divergente 3. Por consideraciones teóricas escogimos una red divergente, pues esta red toma en cuenta la información conocida, es decir, que el tabaquismo produce enfisema y dedos pigmentados y no supone que los investigadores son vírgenes respecto a las hipótesis en juego; en otras redes bayesianas los valores predictivos de la regresión logística y la red bayesiana serían más semejantes (Tabla 3).

Causalidad

La definición de causalidad siempre ha sido motivo de controversia. En este artículo definimos como causa de un efecto a todo factor, condición o característica cuya supresión elimina la posibilidad de que se produzca dicho efecto; y, como factores de riesgo, a las variables asociadas al efecto que, sin ser imprescindibles para que éste se produzca, pueden favorecer la acción de los agentes causales.

Los resultados de la tabla 2 muestran valores estadísticamente significativos para la pigmentación (p=0,001), OR=1,84 [IC 95% 1,27-2,67] pero no para el hábito tabáquico (p=0,366), OR=1,19 [IC 95% 0,82-1,72] en el modelo aditivo. También, bajo un modelo multiplicativo (Tabla 2), la pigmentación es estadísticamente significativa (p=0,018), OR=1,96 [IC 95% 1,12-3,43], y el tabaquismo, no lo es, (p=0,408), OR=1,26 [IC 95% 0,73-2,19]. Entonces, un uso algorítimico y mecánico de la regresión logística conduciría a concluir que es la pigmentación en los dedos y no el tabaquismo la causa del enfisema pulmonar. Mientras para la red bayesiana divergente 3, como veremos enseguida, es todo lo contrario: el tabaquismo y no la pigmentación es la causa de la enfermedad. Las evidencias se presentan en la figura 3.

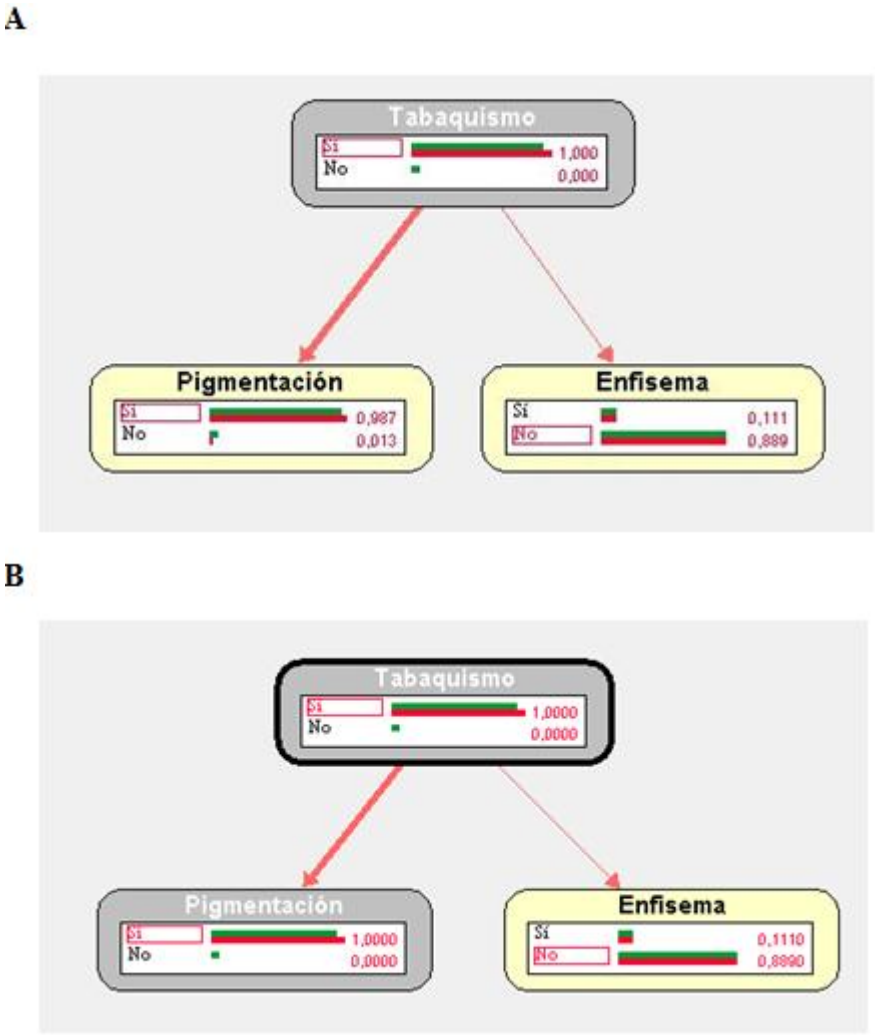


Figura 3. Redes enfisema con evidencia de tabaquismo y con evidencia de tabaquismo más pigmentación.

Los resultados muestran que la probabilidad de tener enfisema, dado que se fuma (Figura 3A), P(+e/+t), tiene el mismo valor que si se sabe que, además, el paciente tiene los dedos pigmentados (Figura 3B), P(+e/+t,+p)=11%. De manera similar, la probabilidad de tener enfisema, dado que no se fuma, P(+e/-t), es igual a si la persona no fuma y tiene los dedos pigmentados, P(+e/-t,+p)=6.1% (tabla 3). Luego, a diferencia de la regresión logística, la red bayesiana divergente 3 da un valor preferente al tabaquismo y no a la pigmentación. Esta red (P←T→E) da lo que se conoce como "d-separación": P U E/T, indicando que si T es conocido, bloquea la comunicación entre P y E (P U E), porque si conocemos que la persona fuma, tener los dedos pigmentados es irrelevante incluso para predecir el enfisema pulmonar. Como afirma, acertadamente, Silva: "Obviamente, si un individuo fumador pudiera eliminar el color amarillo de sus dedos mediante algún tipo de jabón, ello no modificaría en absoluto su riesgo de morir en los próximos diez años debido a que dicha pigmentación no desempeña papel causal alguno en el proceso que se intenta vaticinar" [14].

Un modelo causal final en el cual no intervengan los juicios teóricos de los investigadores no puede obtenerse mediante un método de selección algorítmica, exclusivamente estadístico, de manera automatizada, ya sea de regresión paso a paso en la regresión logística [11] o mediante determinados algoritmos, como K2 y PC en la red bayesiana [15]. Sin embargo, una red bayesiana puede construirse de dos formas: automática y manualmente. Un ejemplo de la primera es tomar los resultados de la tabla 1 y aplicar algoritmos para obtener los enlaces y las probabilidades condicionales que la conforman. En la segunda, con la ayuda de una persona experta en el tema por modelar, construimos un grafo causal (fase cualitativa) y después añadimos las probabilidades condicionales (fase cuantitativa). Las probabilidades pueden obtenerse de una base de datos, de estudios epidemiológicos, literatura médica o estimaciones subjetivas de expertos humanos.

Tampoco debemos descartar una construcción mixta, realizada mediante algoritmos y por una persona experta.

Finalmente, al realizar un razonamiento causal las redes bayesianas permiten tres inferencias. Una es la inferencia predictiva, que va en la dirección de los enlaces desde las causas a los efectos, donde un enlace T→E o T→P es del tipo "T puede causar E" o "T puede causar P". La segunda es una inferencia intercausal, que determina el impacto cualitativo de la evidencia para una variable T sobre una variable P, cuando ambas son antecesoras de una tercera variable E sobre la que existe evidencia independiente (por ejemplo, E ha sido observada), cuya explicación es del tipo "P y T pueden causar E, como T explica E, no existe (casi) evidencia para P" (Figura 4). Por último, una inferencia abductiva que va de los efectos a las causas. Es decir, en el sentido opuesto al de sus enlaces, cuya explicación puede ser "E y P son evidencias para T" (Figura 4B).

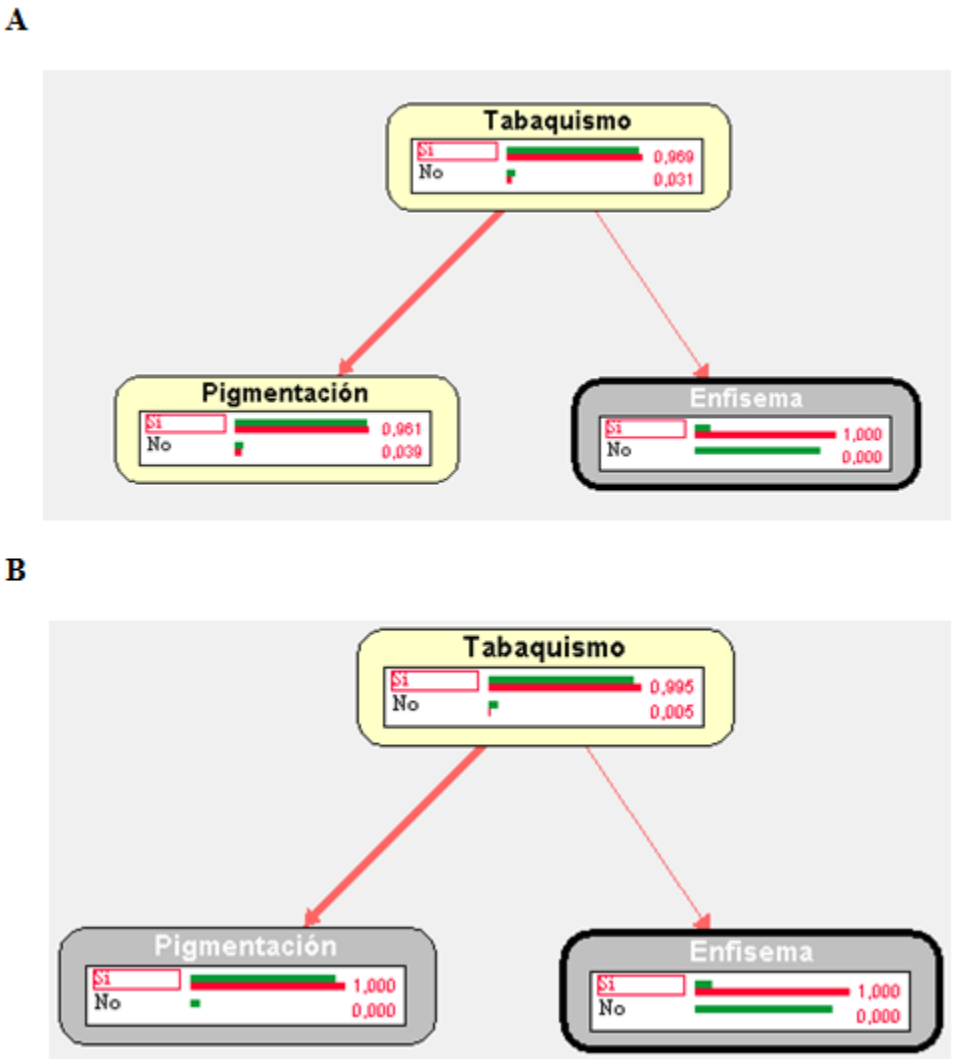


Figura 4. Redes enfisema para predecir el tabaquismo, con dos tipos de evidencias: enfisema y enfisema más pigmentación.

Observemos como, a medida que hay más evidencias (efectos: pigmentación y enfisema), la probabilidad de ser fumador (causa) es mayor: según la tabla 1: P(+t)=0,945 y por las figuras 2, 4A y 4B: P(+t/+p)=0,991, P(+t/+e)=0,969 y P(+t/+p,+e)=0,995. Esta probabilidad inversa se deriva del famoso teorema del reverendo Thomas Bayes sobre "la probabilidad de las causas", publicado en 1764 en *Philosophical Transactions of the Royal Society of London* por su amigo Richard Price.

En este artículo, tenemos una base de datos sobre el enfisema pulmonar, proporcionada por Silva, autor que ha tratado este tema en diferentes artículos y libros [11],[14],[16],[17] y [18], la adaptamos con fines didácticos, y extraemos de ella las variables, e, p y t. De tal

modo que los resultados de la tabla 1 llegaran a las mismas conclusiones que con los datos de Silva en cuanto a la causalidad entre estas tres variables. La pigmentación queda como el factor de riesgo porque presenta un valor p=0,001 y, por no ser relevante el tabaquismo pues presenta un valor p=0,366, queda excluido, (Figura 2). Debido a estos resultados, algunos estadísticos concluirían por medio de la regresión logística que el factor influyente es la pigmentación y no el hábito de fumar, pues "a veces este procedimiento se emplea para descubrir cuáles son las variables causales y desechar por su conducto las que no lo son" [18]. Sin embargo, existe hoy la evidencia de que este, y no aquella, es realmente el que favorece el enfisema pulmonar. Con razón, Silva finaliza: "como es obvio, tal conclusión, lejos de iluminar el camino hacia el conocimiento de las verdaderas relaciones causales, lo ensombrecería" [11]. Porque el procedimiento de regresión paso a paso identifica variables pronósticas, pero nunca conseguirá separar los factores posiblemente causales de los que no lo son [17]. Rothman también, dice "la selección de los factores que entran en un modelo matemático, la formulación matemática del mismo, no son cuestiones de naturaleza estadística; por el contrario, dependen de la comprensión biológica del proceso de enfermar y de las relaciones matemáticas del mismo" [12].

En nuestro ejemplo, el tabaquismo es un factor causal, confusor y predictivo, y la pigmentación es un factor no causal, no confusor pero sí predictivo. Además, aunque en el presente artículo no fue así, debemos tener en cuenta tanto si la introducción de otras variables en la función modifica apreciablemente o no la relación entre la variable dependiente y los factores estudiados, como el margen de error de cada estimación, o su intervalo de confianza, puesto que cualquier medición sin el conocimiento de la incertidumbre carece completamente de significado.

En contraste con la regresión logística, la red bayesiana puede realizar modelos más ricos y ajustados a nuevas evidencias. Otras tres utilidades de la red bayesiana divergente, que merecen consideraciones independientes, y la regresión logística no tiene, son el ajuste de los indicadores de la bondad de una nueva prueba diagnóstica con respecto a una prueba de referencia imperfecta; el aporte a la solución de la famosa paradoja de Simpson, que Judea Pearl (1936) resuelve ya no con tres nodos sino con cinco [19]; y la contribución a la discusión sobre los metaanálisis en red [20].

Conclusión

Este sencillo modelo, aunque no trivial, muestra que las redes bayesianas son una alternativa importante para resolver problemas epidemiológicos que la regresión logística y otras herramientas estadísticas no resuelven. Asimismo, en la toma de decisiones es mejor sumar a una herramienta estadística usada con sentido común, las evidencias anteriores consolidadas durante años o incluso siglos, en lugar del empleo automático y exclusivo de recursos estadísticos.

Notas

Declaración de conflictos de intereses
El autor ha completado el formulario de conflictos de intereses del ICMJE traducido al castellano por*Medwave*, y declara no haber recibido financiamiento para la realización del artículo, y no tener conflictos de intereses asociados a la materia del mismo. Los formularios pueden ser solicitados al autor responsable o a la dirección editorial de la *Revista*.

Referencias

1. Wynder E, Graham E. Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma. A study of six hundred and eighty-four proved cases. JAMA 1950;143: 329-336. |PubMed | Link |
2. Doll R, Hill B. The mortality of doctors in relation to their smoking habits a preliminary report. Br Med J. 1954 Jun;1(4877):1451-5. | PubMed | Link |
3. Fischer RA. Smoking the cancer controversy some attempts to assess the evidence. york.ac.uk [on line] | Link |
4. Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and lung cancer: recent evidence and a discussion of some questions. Int J Epidemiol. 2009 Oct;38(5):1175-91. | CrossRef | PubMed |
5. Berkson J. Smoking and Lung Cancer: Some Observations on Two Recent Reports. J Am Stat Assoc. 1958;53(281):28-38. | Link |
6. Cornfield JT, Gordon T, Smith WW. Quantal response curves for experimentally uncontrolled variables. Bull Int Statist Inst. 1961;38 (3):97-115.
7. Cox D. Commentary: Smoking and lung cancer: reflections on a pioneering paper. Int. J. Epidemiol. 2009;38(5):1192-1193. | Link |
8. Diez FJ. Aplicaciones de los modelos gráficos probabilistas en medicina. Sistemas expertos probabilísticos. Cuenca, España: Universidad de Castilla-La Mancha,1998. | Link |
9. Cooper GF. A diagnostic method that uses causal knowledge and linear programming in the application of Bayes' formula. Comput Methods Programs Biomed. 1986 Apr;22(2):223-37. |PubMed |
10. Lauritzen SL, Spiegelhalter DJ. Local computations with probabilities on graphical structures and their application to expert systems. J Royal Stat Soc. 1988;50(Series B):157-224.
11. Silva LC. La falacia de la regresión de paso a paso. lcsilva.sbhac.net [on line] | Link |
12. Rothman KJ Epidemiología moderna. Madrid, España: Díaz de Santos, 1987.
13. Proyecto Elvira. ia.uned.es [on line] | Link |
14. Silva LC. Causality and prediction: differences and points of contact. Medwave 2014 Sep;14(8):e6016. | CrossRef | PubMed |
15. Morales M, Salmerón A, Rodríguez C. Análisis de indicadores de rendimiento mediante redes bayesianas. España: Editorial Universidad de Almería, 2007.
16. Silva LC. Regresión logística. Madrid, España: La muralla. S.A, Hespérides S.A, 2004.

17. Silva LC. Adocenamiento y ceremonias metodológicas. En: La investigación biomédica y sus laberintos. En defensa de la racionalidad para la ciencia del siglo XXI. Madrid, España: Díaz de Santos, 2009:293-330.
18. Silva LC. Estudios de casos y controles en psiquiatría: causalidad, diseño y advertencias. Actas Esp Psiquiatr 2004;32(4):236-248. | Link |
19. Pearl J. Simpson's Paradox, confounding, and collapsibility. En: Causality. Models, reasoning and inference. Cambridge: University Press, 2013:293-330.
20. Catalá-López F, Tobías Aurelio, Roqué Marta. Conceptos básicos del metaanálisis en red. Aten Primaria. 2014;46(10):573-581. | Link |

Correspondencia a:
Carrera 98 N° 44-33,
Apartamento 201,
Medellín,
Colombia