Ronda clínica y epidemiológica. Introducción al análisis multivariable (parte II)

Daniela Garcés[1], Fabián Jaimes Barragán[2]

RESUMEN

En la primera parte de este tema presentamos su definición, principales usos y los tres métodos de análisis multivariable más utilizados en la literatura científica. En esta segunda parte profundizaremos en los criterios para la incorporación de variables independientes al análisis, las herramientas para evaluar qué tan adecuado es el modelo seleccionado y la interpretación de los resultados y de los coeficientes en cada tipo de regresión.

PALABRAS CLAVE

Análisis Multivariable; Epidemiología; Regresión Lineal; Regresión Logística; Regresión de Cox

SUMMARY

Clinical and epidemiological round. Introduction to multivariable analysis (part II)

In the first part of this topic, we explained the definition of multivariable analysis, its main uses, and the three most commonly used methods in the scientific literature. In this second part, we will delve into the inclusion criteria of independent variables to the analysis, the tools to assess how the model fits the data and the interpretation of the results and coefficients in each regression model.

KEY WORDS

Epidemiology; Cox Models; Linear Models; Logistic Models; Multivariable Analysis

RESUMO

Ronda clínica e epidemiológica. Introdução à análise multi-variáveis (parte II)

Na primeira parte deste tema apresentamos sua definição, principais usos e os três métodos de análises multi-variável mais utilizados na literatura científica. Nesta segunda parte aprofundaremos nos critérios para a incorporação de variáveis independentes à análise, as ferramentas

[1] Estudiante de Medicina, Joven investigadora, Grupo Académico de Epidemiología Clínica (GRAEPIC), Universidad de Antioquia, Medellín, Colombia.
[2] Profesor Titular, Grupo Académico de Epidemiología Clínica (GRAEPIC), Departamento de Medicina Interna, Facultad de Medicina, Universidad de Antioquia, e Investigador, Unidad de Investigaciones, Hospital Pablo Tobón Uribe, Medellín, Colombia.
Correspondencia: Fabián Jaimes Barragán; fabian. jaimes@udea.edu.co

Recibido: septiembre 08 de 2014
Aceptado: septiembre 25 de 2014

para avaliar que tão adequado é o modelo selecionado e a interpretação dos resultados e dos coeficientes em cada tipo de regressão.

PALAVRAS CHAVE

Análises Multi-variável; Epidemiologia; Regressão Linear; Regressão Logística; Regressão de Cox

INCORPORACIÓN DE VARIABLES INDEPENDIENTES EN EL ANÁLISIS MULTIVARIABLE

Las variables independientes dicotómicas tienen una ventaja con respecto a otros tipos de variables como las ordinales, las nominales y de intervalo, ya que pueden ser analizadas con cualquier método estadístico sin necesidad de ninguna transformación más allá de la simple recodificación como 1 (presencia de la variable) o 0 (ausencia de la misma) (1). Con los otros tipos de variables, en cambio, es necesario efectuar transformaciones específicas para poder analizarlas. Las variables nominales representan atributos no numéricos que no admiten un tipo de orden, por lo que al recibir códigos numéricos para efectos de nomenclatura, estos no tienen ningún significado (2). Por ejemplo, cuando vemos el estado civil de una persona como soltero, casado, viudo o divorciado, cada una de esas alternativas para la variable recibe un código numérico como soltero = 1, casado = 2, viudo = 3, divorciado = 4. Estos números, no obstante, no reflejan ninguna magnitud y en cualquier análisis multivariable el pasar de una categoría a otra, de 1 a 2 por ejemplo, no tendría ningún significado. Para que las variables nominales independientes puedan tener un valor "real" y ser analizadas, deben convertirse en múltiples variables dicotómicas por medio de un proceso que en inglés se ha llamado *dummying* (3). Si tomamos el ejemplo anterior y suponemos que cada una de las opciones de estado civil es una variable dicotómica, estas recibirían el valor de 1 para indicar que el individuo está en esa categoría y 0 si no lo está. Las nuevas variables dicotómicas quedarían casado (1 = sí, 0 = no), viudo (1 = sí, 0 = no), divorciado (1 = sí, 0 = no) y soltero (1 = sí, 0 = no). Dado que realmente se están representando cuatro categorías excluyentes de una misma variable, este proceso de transformación siempre genera por defecto el valor de la última variable. Es decir, si un sujeto está codificado como no casado, no viudo y no divorciado (con valor de 0 en las primeras tres variables) la única opción que le queda es ser soltero, es decir código de 1 en la última variable. Por lo anterior, en este proceso de *dummying*, el investigador siempre decide cuál de las opciones de la variable se convertirá en la categoría de referencia contra la cual se comparan todas las otras alternativas; y a este conjunto de categorías se le llama variable *dummy* o indicadora (3). Todas las anteriores consideraciones aplican por igual para las variables ordinales, con la diferencia que en estas sí hay un orden jerárquico entre las diversas opciones de la variable y de este modo la categoría de referencia está seleccionada de manera natural.

Crane y colaboradores (4) hicieron una investigación para buscar la asociación entre los altos niveles de glucosa en pacientes diabéticos y no diabéticos y el riesgo de desarrollar demencia. En la tabla 1 se muestran dos ejemplos de variables nominales de ese estudio.

Tabla 1. Ejemplos de variables nominales para un estudio clínico

	Características					Diabetes n = 232	No diabetes n = 1.835	OR
Raza	Blanco	*1*	*0*	*0*	*0*	190	1.673	1 (ref)
	Negro	*0*	*1*	*0*	*0*	28	72	3,42
	Asiático	*0*	*0*	*1*	*0*	9	58	1,36
	Otro	*0*	*0*	*0*	*1*	5	32	1,37
Fumar	Fumador activo		*1*	*0*	*0*	6	98	1 (ref)
	Exfumador		*0*	*1*	*0*	125	853	2,39
	Nunca ha fumado		*0*	*0*	*1*	101	884	1,80

Adaptación de Crane K. Paul et al. Glucose levels and risk of dementia. N Engl J Med 2013; 369:540-8

Tanto la raza como el fumar son variables nominales que tomaron un valor dicotómico para ser comparadas entre diabéticos y no diabéticos. De 232 diabéticos, 190 están codificados como 1 en raza blanca y 0 en negros, asiáticos y otros; 28 están como 1 en raza negra y 0 en blancos, asiáticos y otros; 9 aparecen como 1 en asiáticos y 0 en blanco, negro y otros; y el 0 de las tres categorías anteriores obliga a tener un valor de 1 en la raza "otro" que corresponde a 5 sujetos. Las mismas consideraciones, con sus respectivos números, aplican para la raza en no diabéticos y el fumar en diabetes y no diabetes. No hay un esquema único de cómo agrupar las variables nominales ni de la categoría de referencia que se debe seleccionar para comparar. La mejor forma de hacerlo dependerá de la pregunta de investigación, de la frecuencia y distribución de las variables y sus categorías, y de la misma asociación estimada o sospechada entre las variables independientes y la variable de desenlace (1). En este ejemplo, la raza blanca por ser la de mayor población o el fumar por ser un conocido factor de riesgo podrían ser tomados como categorías de referencia y asignarles un valor de cero para las comparaciones finales. El *Odds Ratio* (OR), o la medida de asociación necesaria de acuerdo con el modelo, se interpreta en comparación con la categoría de referencia seleccionada. En el ejemplo anterior, la raza negra tiene una asociación 3,42 veces mayor que la raza blanca con la presencia de diabetes; y los asiáticos, a su vez, una asociación 36% mayor que los blancos con respecto al mismo desenlace.

En cuanto a las variables continuas o de intervalo, el modelo multivariable asume que cada cambio de una unidad, en cualquier punto de la escala de la variable independiente, tiene un cambio de igual magnitud en la variable de resultado del modelo (asunción de linealidad) (2). Por ejemplo, si la variable independiente es edad, un aumento de la edad de 1 año genera igual cambio si es de 30 a 31 o de 80 a 81 años. Por supuesto, la medida de este cambio será diferente dependiendo de la variable de resultado y por tanto del método de análisis multivariable que se use: lineal, logística o Cox (1).

En la regresión lineal este supuesto de linealidad se aprecia fácilmente en un diagrama de dispersión que grafica en el plano cartesiano la variable independiente como el eje de las "x" y la variable dependiente como el eje de las "y" (figura 1). Pero en el caso de la regresión logística y la regresión de Cox, por la transformación que necesitan sus respectivas variables de resultado, no es posible hacer un diagrama de dispersión convencional. Una aproximación para poder evaluar si una variable de intervalo se ajusta a la hipótesis de linealidad en estos casos es categorizarla en múltiples variables dicotómicas que contengan las mismas unidades en la escala de la variable (2). Por ejemplo, si tomamos como variable independiente la edad, entonces la agruparíamos de 10 en 10 así: 20-29, 30-39, 40-49, 50-59 años, y le daríamos el valor de 1 o 0 a cada individuo de la muestra, dependiendo de si pertenece (1) o no (0) al grupo correspondiente. Luego de este proceso, las nuevas variables deben entrar a un análisis multivariable en donde reciben un coeficiente estimado de acuerdo con su respectivo efecto de incremento o disminución de la variable de resultado (1). Este coeficiente podría graficarse en el eje de las "y" correspondiente con el punto medio de cada grupo marcado en el eje de las "x" y verificar si el resultado se asemeja a la línea recta que se esperaría en un diagrama de dispersión (5).

Ocasionalmente, estos modelos gráficos pueden sugerir que la variable independiente continua se relaciona con el desenlace, pero no de manera lineal sino de alguna otra forma como exponencial, logarítmica, curvilínea o con un efecto umbral. En estos casos, existen tres métodos que permiten la modificación e inclusión de estas variables al modelo de análisis multivariable: a) las transformaciones matemáticas, la más usada de las cuales es la transformación logarítmica; b) los *splines*, que son polinomiales o funciones que al sumar términos algebraicos pueden "conectar" los segmentos de una línea irregular; y finalmente, c) la conversión en múltiples variables dicotómicas o creación de variables indicadoras como se explicó previamente (1). Infortunadamente, estos tres métodos pueden presentar inconvenientes como la dificultad en la interpretación de sus resultados, los distintos cálculos matemáticos que requiere su elaboración y la escogencia de puntos de corte arbitrarios en el caso de las variables *dummy*.

INCLUSIÓN Y EXCLUSIÓN DE LAS VARIABLES INDEPENDIENTES

Para incluir las variables independientes en el análisis multivariable que busca resolver una pregunta de

etiología o causalidad, se deben tener en cuenta todos los aspectos relacionados con la pregunta de investigación y cualquier variable que potencialmente pueda afectar la relación entre la exposición y el desenlace. Estas variables, llamadas de confusión, distorsionan la medida de asociación entre la exposición y el desenlace dando como resultado la observación de un efecto que en realidad no existe, la exageración o la atenuación de una asociación real (6). Para incluir este tipo de variables se recomienda escoger aquellas que ya han sido teorizadas o se mostraron como una variable de confusión en investigaciones anteriores. También se pueden incluir variables que empíricamente, es decir, en el mismo análisis de los resultados actuales, se definan como factores de confusión con base en detectar su asociación con el factor de riesgo/exposición y también con el desenlace. Algunos investigadores incluyen, a pesar de ser una estrategia bastante discutible, cualquier variable que presente una asociación con el desenlace y tenga una $p < 0{,}20$-$0{,}25$, independientemente de si se ha demostrado con anterioridad cualquier asociación con dicha variable (1).

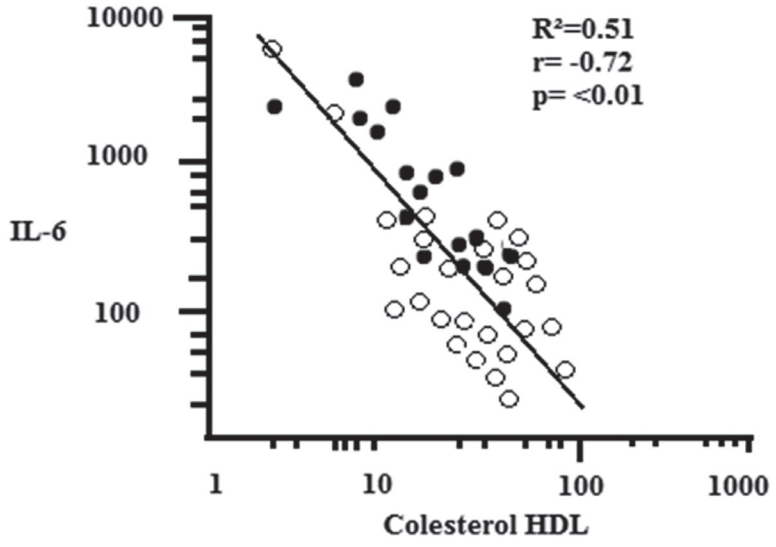

Figura 1. Diagrama de dispersión para la correlación entre los niveles de interleucina 6 (IL-6) y colesterol HDL en sepsis. Adaptación de Jung-Yien Chien. Low serum level of high-density lipoprotein cholesterol is a poor prognostic factor for severe sepsis. Crit Care Med 2005; 33 (8): 1688-93

Es importante excluir las variables que definitivamente no están en la vía causal que se está analizando, las variables redundantes (ver adelante multicolinealidad) y las variables denominadas intervinientes. Estas últimas son las que se encuentran en la vía causal del desenlace, pero son desencadenadas o causadas por el mismo factor de riesgo en estudio (1). Si en el análisis multivariable se ajusta por una variable interviniente, el factor de riesgo en estudio perdería valor en la asociación con el desenlace. Por ejemplo, el consumo moderado de alcohol está asociado con una menor incidencia de enfermedad coronaria. En este caso, la variable independiente "alcohol" disminuye el riesgo de enfermedad coronaria debido al aumento que produce el consumo moderado del mismo en el colesterol HDL. Sin embargo, si se ajusta por la variable "HDL" el alcohol perdería valor en la asociación, ya que el aumento del HDL es la causa fundamental de la disminución del riesgo cardiovascular (1).

MULTICOLINEALIDAD

La multicolinealidad se produce cuando dos o más variables independientes son tan estrechamente relacionadas una con la otra, que el modelo puede no ser capaz de evaluar de forma fiable la contribución independiente de cada una de ellas (2). La multicolinealidad implica la existencia de una relación lineal "perfecta o exacta" entre algunas de las variables independientes de un modelo de regresión (2). Si se incluyen dos o más variables multicolineales en un modelo de regresión los coeficientes respectivos no serán confiables, ya que tendrán grandes distorsiones en la dirección del efecto y mayores errores estándar, lo que le impedirá al lector juzgar la precisión de los resultados. El diagnóstico final de la multicolinealidad requiere la realización de una medida denominada factor de inflación de la varianza (VIF), la cual mide qué tanto del coeficiente de regresión de una variable es determinado por las otras variables independientes del modelo (1). Cuando se encuentra la multicolinealidad es posible tomar alguna de las siguientes opciones: omitir la variable que es teóricamente menos importante, la que presenta más valores faltantes o de alguna manera es menos satisfactoria para el análisis, o crear nuevas combinaciones de variables con diversas categorías o con escalas más elaboradas.

DIAGNÓSTICO DE LOS MODELOS E INTERPRETACIÓN DE SUS RESULTADOS

Todos los tipos de análisis multivariable que conocemos en investigación clínica tratan de representar un conjunto de datos (es decir, las variables de todos los individuos de la población de estudio) por medio de un modelo estadístico. Por lo tanto, el primer paso antes de evaluar e interpretar los resultados del análisis es determinar qué tan correcta es la representación que hace el modelo estadístico de esos datos. Dicha evaluación se hace por medio de pruebas estadísticas, usualmente denominadas "pruebas de bondad de ajuste", que varían de acuerdo con el tipo de análisis multivariable que se considere.

REGRESIÓN LINEAL MÚLTIPLE

El coeficiente de variación o R^2 mide la cantidad de la variabilidad de la variable dependiente que es explicada por las variables independientes incluidas en el modelo (7). En otras palabras, es una medida cuantitativa del ajuste de los datos con un valor entre 0 y 1; el valor 0 indica que definitivamente las variables independientes no explican el resultado y el valor 1 indica que las variables explican perfectamente el desenlace. Una limitación importante del R^2 es que a medida que se introducen más variables independientes en el modelo hace una estimación de la importancia relativa que tiene cada variable nueva; lo que quiere decir que el R^2 siempre aumentará al incluir variables, aunque estas no sean realmente significativas para predecir el cambio en la variable dependiente (7). Por lo anterior, la verificación de lo adecuado del modelo se debe hacer con el "R^2 ajustado", que tiene en cuenta también el número de variables que se incluyen en la regresión y "penaliza" el exceso de las mismas. En la regresión lineal múltiple se modela el valor promedio del desenlace. Es decir, por cada unidad de aumento de la variable independiente aumenta o disminuye, en la cantidad que indica el coeficiente, el promedio de la variable dependiente. Un coeficiente positivo indica que cada aumento de la variable independiente se acompaña también de un aumento en la variable dependiente; y un coeficiente negativo indica que el aumento de la primera, por el contrario, se acompaña de un descenso en la segunda. Este coeficiente, si se "superpone" en un diagrama de dispersión como el del ejemplo anterior (figura 1), corresponde a la pendiente o inclinación en la línea que mejor representa la relación entre la variable independiente y el desenlace (1). Para cada uno de los coeficientes que se obtienen en la regresión lineal múltiple el programa estadístico calcula el valor de p, que es la probabilidad de obtener resultados iguales o más extremos que los observados asumiendo que la hipótesis nula es correcta (8). Si este valor de p es pequeño, usualmente $< 0{,}05$, se considera estadísticamente significativo y se rechaza la hipótesis nula de que el coeficiente es igual a 0, o en otras palabras, de que no existe pendiente en la línea.

Martin Senechal y colaboradores (9) hicieron una investigación para determinar si una mejor aptitud cardiorrespiratoria atenúa los defectos estructurales y funcionales del ventrículo izquierdo y la esteatosis miocárdica en pacientes con diabetes mellitus tipo 2. Las variables de resultado fueron la estructura y la función del ventrículo izquierdo (medidas con fracción

de eyección y volumen de fin de diástole) y la cantidad de triglicéridos (TG) en el corazón medidos con espectroscopía de resonancia magnética, la cual da como resultado una proporción de grasa/agua (%). La variable principal de exposición fue definida por el consumo máximo de oxígeno para la masa libre de grasa (VO2 peak-FFM) y los pacientes fueron divididos en terciles de acuerdo con dicha cantidad. El primer tercil tenía un valor de 17,85-25,93; el segundo de 25,94-30,14; y el tercero, que era el grupo de referencia, tenía un VO2 peak-FFM de 30,15-43,97. Se utilizó la regresión lineal múltiple para evaluar la anterior asociación y se ajustó por edad, sexo, raza, tabaquismo, presión arterial sistólica, dislipidemia, índice de masa corporal, duración de la diabetes y el uso de insulina. Los resultados se muestran en la tabla 2.

Tabla 2. Relación entre la cantidad de TG en el corazón, la función miocárdica y la aptitud cardiorrespiratoria por medio de un modelo de regresión lineal

	Esteatosis miocárdica (%), valor p	VFDVI (mL), valor p	Fracción de eyección (%), valor p
Tercil 1	0,241 (0,343)	-14,85 (0,032)	0,174 (0,952)
Tercil 2	0,487 (0,074)	-5,95 (0,402)	-1,676 (0,587)
Tercil 3	Referencia	Referencia	Referencia

TG: triglicéridos. VFDVI: volumen final de diástole del ventrículo izquierdo. Tercil 1: VO2 peak-FFM 17,85-25,93. Tercil 2: VO2 peak-FFM 25,94-30,14. Tercil 3: VO2 peak-FFM 30,15-43,97. Adaptación de Senechal M et al. Is cardiorespiratory fitness a determinant of cardiomyopathy in the setting of type 2 diabetes? Diabetes and Vascular Disease Research 2014; 11(5):343-51.

En la tabla 2 se puede observar que los individuos del tercil 1 de consumo de oxígeno tuvieron en promedio 0,241% más de TG en el corazón que los pacientes del tercil 3, pero esta diferencia no fue estadísticamente significativa (p = 0,343). Con respecto al desenlace de VFDVI, en cambio, se aprecia que los sujetos del primer tercil tienen en promedio 14,85 mL menos que los del tercil 3, y esta diferencia sí alcanzó significado estadístico (p = 0,032). En el tercil 2 de consumo de oxígeno, si bien también hay un menor valor promedio de VFDVI con respecto al tercil de referencia (-5,95), la diferencia no alcanzó significación (p = 0,402).

REGRESIÓN LOGÍSTICA MÚLTIPLE

Para evaluar qué tan bien se ajusta el modelo de regresión logística a los datos se utiliza una prueba estadística de verosimilitud (en inglés: *likelihood ratio test)*, que determina si las variables independientes se asocian con el desenlace de interés más de lo que podría esperarse solo por azar. Cuando esta prueba muestra que la proporción de sujetos con el desenlace se puede explicar por la forma como se combinan las variables independientes del modelo, su resultado se acompaña de un número alto en la distribución de Chi^2 y por tanto de un menor valor de la p correspondiente. Un valor de $p < 0,05$, al igual que en la regresión lineal, rechaza la hipótesis nula de que no existe asociación entre las variables independientes y el resultado (1).

En la regresión logística el significado de los coeficientes es diferente al que encontramos en la regresión lineal ya que en estos casos lo que se modela es el logaritmo del *odds* (logit), como explicamos en el número anterior de esta serie (10). Un coeficiente positivo quiere decir que a medida que aumenta la variable independiente aumenta el logit, y un coeficiente negativo indica que a medida que aumenta la variable independiente el logit disminuye. El paquete estadístico arroja un OR *(Odds Ratio)* para cada variable independiente, luego de ajustar de manera simultánea por todos los factores presentes en el modelo, que es igual al antilogaritmo del coeficiente (fórmula 1) e indica qué tanto aumenta (OR > 1) o disminuye (OR < 1) el riesgo de ocurrencia del desenlace por cada cambio en una unidad en la variable independiente (11).

Fórmula 1

Odds ratio = $e^{coeficiente}$

Al igual que en la regresión lineal, es posible determinar el significado estadístico de cada coeficiente de la regresión logística por medio del valor de p que se obtiene con la prueba de Wald (12), también rechazando la hipótesis nula de no asociación con un valor $p < 0,05$. Sin embargo, es preferible y mucho más ilustrativo interpretar los OR con base en la precisión que se observa con los intervalos de confianza del 95% (IC 95%).

Carreno y colaboradores (13) hicieron una investigación para evaluar la incidencia y los factores de riesgo para desarrollar insuficiencia renal aguda (IRA) en pacientes en tratamiento con vancomicina. La población de estudio se dividió en tres categorías según la edad: adultos jóvenes (de 18 a 64 años), adultos mayores (de 65 a 79 años) y ancianos (80 años o más). Se utilizó la regresión logística para cuantificar el efecto simultáneo de las siguientes variables en el riesgo de ocurrencia de IRA: grupo de edad, infección del tracto respiratorio inferior, duración de la terapia en días y presencia de al menos dos factores de riesgo conocidos para nefrotoxicidad. Los resultados se muestran en la tabla 3.

Tabla 3. Factores de riesgo para desarrollar insuficiencia renal aguda en pacientes tratados con vancomicina (modelo de regresión logística múltiple)

Variables	OR (IC 95%)	p
Adultos jóvenes	1,00 (referencia)	
Adultos mayores	0,69 (0,25-1,92)	0,48
Ancianos	0,78 (0,28-2,26)	0,80
Infección del tracto respiratorio inferior	5,18 (2,15-12,4)	< 0,01
Duración del tratamiento (días)	1,12 (1,03-1,22)	< 0,01
Dos o más factores de riesgo para nefrotoxicidad	6,94 (1,81-26,7)	< 0,01

Adaptación de Carreno JJ. Comparative incidence of nephrotoxicity by age group among adult patients receiving vancomycin. Infect Dis Ther 2013; 2(2): 201-208.

Los resultados obtenidos en la regresión logística sugieren que los grupos de mayor edad, comparados con los jóvenes, parecerían tener menos riesgo de IRA con el uso de vancomicina. Sin embargo, esta aparente disminución del riesgo no es estadísticamente confiable como indican los límites del intervalo de confianza y los valores de p respectivos. Por otra parte, tener una infección del tracto respiratorio inferior, comparada con otros tipos de infección como indicaciones para el uso de vancomicina, incrementa en más de 5 veces el riesgo de desarrollar IRA, independientemente del grupo de edad, la duración del tratamiento y la presencia o ausencia de factores de riesgo para nefrotoxicidad. Este aumento del riesgo puede variar entre 2,15 y 12,4, pero es constante en su comportamiento por encima del 1. Del mismo modo, cada día de tratamiento aumenta un 12% el riesgo de IRA en cualquier grupo de edad, en cualquier infección de base y en presencia o ausencia de otra nefrotoxicidad, con una variación del aumento de riesgo entre 3% y 22%.

REGRESIÓN DE COX

La regresión de Cox se utiliza cuando la variable dependiente describe el tiempo transcurrido hasta la ocurrencia de un determinado evento en un conjunto de individuos, y se quiere determinar

simultáneamente el efecto independiente de una serie de factores que pueden alterar dicho tiempo. Frecuentemente, el evento de interés es la mortalidad y por esta razón estos modelos también se conocen como análisis de supervivencia. A diferencia de otros tipos de análisis, en estos es necesario incluir la variable del tiempo para cada uno de los sujetos participantes. Este tiempo se determina para cada individuo como el intervalo desde que ingresa al análisis hasta que presenta el evento de interés o hasta que se completa el tiempo de seguimiento estipulado inicialmente en la investigación, así no se haya presentado aún el desenlace. Además, en los sujetos que se pierden del seguimiento antes de lo previsto también se registra el respectivo tiempo de observación (1).

Es necesario escoger el punto de partida, o el ingreso al análisis, que mejor represente el inicio del proceso que se está estudiando, y es de vital importancia aclarar el porqué de su elección: el momento de la asignación aleatoria en los estudios experimentales, o la primera visita médica, el primer síntoma o el momento del diagnóstico, entre otros, en los estudios observacionales. La unidad de medida de tiempo seleccionada para el análisis depende fundamentalmente del curso clínico de la enfermedad: en días e incluso en horas para las enfermedades de evolución rápida y en años para las de progresión lenta.

Al igual que en la regresión logística, las pruebas que miden qué tanto el modelo de regresión de Cox se ajusta a los datos y qué tan significativos son los coeficientes obtenidos son el *likelihood ratio test* y la prueba de Wald, respectivamente (12). El primero determina qué tanto el conjunto de variables independientes incorporadas en el modelo se asocia con el resultado más de lo esperado por el azar, y el segundo indica la significación estadística del coeficiente correspondiente a cada una de dichas variables. La interpretación de los valores de p resultantes de las anteriores pruebas, y la preferencia por el uso de los intervalos de confianza para conocer la precisión de la medida de asociación, sigue los mismos lineamientos que se explicaron para la regresión logística.

La regresión de Cox modela el logaritmo del HR (*hazard* relativo), que representa la magnitud en la que una variable aumenta o disminuye el riesgo (el *hazard* del inglés) de ocurrencia de un desenlace en el tiempo. Por tanto, el coeficiente indica cuánto cambio hay en el logaritmo del HR por cada cambio de una unidad en la variable independiente. El significado de los signos del coeficiente es similar al de la regresión logística: si un coeficiente es positivo indica que por cada aumento de una unidad en la variable independiente aumenta el logaritmo del HR, y un coeficiente negativo indica que por cada aumento de la variable independiente disminuye el logaritmo del HR. Al igual que el OR en la regresión logística, el HR corresponde al antilogaritmo del coeficiente obtenido (fórmula 2) y se interpreta de manera similar. Por tanto, un HR igual a 1 quiere decir que no hay ningún cambio en el tiempo transcurrido hasta un evento a pesar de observar cambios en la variable independiente. Un HR mayor de 1 indica que cuando la variable independiente está presente aumenta o "acelera" el riesgo de tener un evento en un tiempo determinado; y cuando el HR es menor de 1 la presencia o el aumento de dicha variable independiente disminuye o "frena" el riesgo de aparición del desenlace.

Fórmula 2

$$HR = e^{\text{coeficiente}}$$

Clague y colaboradores (14) hicieron una investigación para examinar la asociación entre el uso de la terapia hormonal en la menopausia (THM) y la mortalidad específica por cáncer de pulmón en una cohorte de 727 mujeres reclutadas entre 1995-1996 y seguidas hasta 2007. El punto de partida para el seguimiento fue el momento del diagnóstico de cáncer de pulmón y los puntos finales para el seguimiento fueron la muerte, mudarse de los Estados Unidos, o diciembre 31 de 2007. Se clasificó la población en quienes nunca habían usado THM y quienes sí la habían usado o la seguían usando, además del tiempo de uso: menos de 5 años, entre 6 y 15 años o más de 15 años. Se usó la regresión de Cox para el análisis multivariable ajustando por edad, raza, fumador reciente, exfumador, fumador activo y estado del tumor. Los resultados se muestran en la tabla 4.

Tabla 4. Asociación entre la terapia de reemplazo hormonal con estrógenos y la mortalidad en mujeres con cáncer de pulmón

	Total muertes/mujeres	HR 95% IC
Nunca	144/216	Referencia
Quienes la usaron o la siguen usando	155/254	0,69 (0,52-0,93)
Duración		
<5 años	54/87	0,59 (0,39-0,87)
6 y 15 años	37/58	0,84 (0,52-1,35)
>15 años	55/95	0,63 (0,41-0,96)

Adaptación de Clague J et al. Menopausal hormone therapy and lung cancer-specific mortality following diagnosis: The California Teachers Study. PLoS ONE 2014; 9(7): e103735014.

Se observa que hubo una disminución en el riesgo *(hazard)* de muerte por cáncer de pulmón en el grupo que usó THM comparado con el grupo que nunca la usó (HR = 0,69; IC 95% = 0,52-0,93). Este HR indica que el uso de THM, independientemente de (o ajustado por) la edad, la raza, la condición de fumar o el estado del tumor de las mujeres de la cohorte, disminuye en un 31% (1-0,69) el riesgo de morir por cáncer de pulmón en mujeres menopáusicas durante los años posteriores al diagnóstico de cáncer. Dicha disminución, de acuerdo con el intervalo de confianza respectivo, podría ser de 48% (1-0,52) o de 7% (1-0,93), pero siempre indicando que el uso de THM parece asociarse con un mejor pronóstico en estos casos. El tiempo de uso de la THM, sin tener en cuenta la significación estadística de los respectivos HR, no parece modificar los anteriores hallazgos y su interpretación.

CONCLUSIÓN

Para hacer un uso correcto de los resultados del análisis multivariable es necesario que el lector esté familiarizado con el papel que cumplen las variables independientes y con su adecuada codificación y/o inclusión en el modelo seleccionado. Además de ciertos supuestos o asunciones básicas de la regresión, se deben tener presentes los métodos para hacer la evaluación del ajuste de los modelos y la interpretación de sus respectivos resultados.

REFERENCIAS BIBLIOGRÁFICAS

1. Katz MH. Multivariable analysis: a practical guide for clinicians and public health researchers. 3rd ed. Cambridge: Cambridge University Press; 2011.
2. Kelmansky DM. Estadística para todos: estrategia de pensamiento y herramientas para la solución de problemas. Buenos Aires: MInisterio de Educación de la Nación; 2009.
3. Szklo M, Nieto J. Epidemiology: beyond the basics. Hudson MA: Jones and Bartlett Books; 2006.
4. Crane PK, Walker R, Hubbard RA, Li G, Nathan DM, Zheng H, et al. Glucose levels and risk of dementia. N Engl J Med. 2013 Aug 8;369(6):540–8.
5. Chien J-Y, Jerng J-S, Yu C-J, Yang P-C. Low serum level of high-density lipoprotein cholesterol is a poor prognostic factor for severe sepsis. Crit Care Med. 2005 Aug;33(8):1688–93.
6. de Irala J, Martínez-González MA, Guillén Grima F. [What is a confounding variable?]. Med Clin (Barc). 2001 Oct 6;117(10):377–85.
7. Rodríguez Jaume MJ, Mora Catalá R. Análisis de regresión múltiple. Alicante: Universidad de Alicante; 2001.
8. Manterola D C, Pineda N V. El valor de "p" y la "significación estadística": Aspectos generales y su valor en la práctica clínica. Rev Chil Cir. 2008 Feb;60(1):86–9.

9. Sénéchal M, Ayers CR, Szczepaniak LS, Gore MO, See R, Abdullah SM, et al. Is cardiorespiratory fitness a determinant of cardiomyopathy in the setting of type 2 diabetes? Diab Vasc Dis Res. 2014 Sep;11(5):343–51.

10. Garcés D, Jaimes Barragán F. Ronda clínica y epidemiológica: Introducción al análisis multivariable (parte I). Iatreia. 2014;27(3):355–63.

11. Fiuza MD, Rodríguez Pérez JC. [Logistic regression: a versatile tool]. Nefrologia. 2000;20(6):495–500.

12. Fernandez M, Abraira V, Quereda C, Ortuño J. Curvas de supervivencia y modelos de regresión: errores y aciertos en la metodología de aplicación. Nefrologia. 1996;XVI(5):383–90.

13. Carreno JJ, Jaworski A, Kenney RM, Davis SL. Comparative Incidence of Nephrotoxicity by Age Group among Adult Patients Receiving Vancomycin. Infect Dis Ther. 2013 Dec;2(2):201–8.

14. Clague J, Reynolds P, Henderson KD, Sullivan-Halley J, Ma H, Lacey J V, et al. Menopausal hormone therapy and lung cancer-specific mortality following diagnosis: the California Teachers Study. PLoS One. 2014 Jan;9(7):e103735.