Colegio Mexicano de Inmunología Clínica y Alergia, A. C. | CMICA
Revista Alergia México raM

Metodología de la investigación

Correlation: not all correlation entails causality

Correlación: no toda correlación implica causalidad

Ivonne Roy-García,[1] Rodolfo Rivas-Ruiz,[1] Marcela Pérez-Rodríguez,[1] Lino Palacios-Cruz[2]

Abstract

The concept of correlation entails having a couple of observations (X and Y), that is to say, the value that Y acquires for a determined value of X; the correlation makes it possible to examine the trend of two variables to be grouped together. We know that, with increasing age, blood pressure figures also increase, therefore, if we want to answer a research question like "what is the connection between age and blood pressure?" the relevant statistical test is a correlation test. This test makes it possible to quantify the magnitude of the correlation between two variables, but it is also helpful for predicting values. If these variables had a perfect correlation, the value of the variable Y could be deduced by knowing the value of X. Because of these advantages, the correlation is one of the most frequently used tests in the clinical setting since, in addition to measuring the direction and magnitude of the association of two variables, it is one of the foundations for prediction models, such as linear regression model, logistic regression model and Cox proportional hazards model.

Keywords: Clinical research; Prediction models; Statistical correlation

Este artículo debe citarse como: Roy-García I, Rivas-Ruiz R, Pérez-Rodríguez Marcela, Palacios-Cruz L. Correlación: no toda correlación implica causalidad. Rev Alerg Mex. 2019;66(3):354-360

ORCID
Ivonne Roy-García, 0000-0002-1859-3866; Rodolfo Rivas-Ruiz, 0000-0002-5967-7222;
Marcela Pérez-Rodríguez, 0000-0003-3417-3201; Lino Palacios-Cruz, 0000-0001-9533-2996

[1]Instituto Mexicano del Seguro Social, Centro Médico Nacional Siglo XXI, Centro de Adiestramiento en Investigación Clínica, Ciudad de México, México
[2]Instituto Nacional de Psiquiatría Dr. Ramón de la Fuente, Subdirección de Investigaciones Clínicas, Ciudad de México, México

Correspondencia: Rodolfo Rivas-Ruiz. rivasrodolfo@gmail.com

Recibido: 2019-07-30
Aceptado: 2019-08-01
DOI: 10.29262/ram.v66i3.651

Resumen

El concepto de correlación implica contar con un par de observaciones (X y Y), es decir, el valor que toma Y para determinado valor de X; la correlación permite examinar la tendencia de dos variables a ir juntas, por ejemplo, sabemos que al incrementar la edad también aumentan las cifras de presión arterial, por lo tanto, si queremos responder una pregunta de investigación como ¿cuál es la relación entre edad y presión arterial?, la prueba estadística pertinente es una prueba de correlación. Esta prueba permite cuantificar la magnitud de la correlación entre dos variables y ayuda a predecir valores. Si estas variables tuvieran una correlación perfecta se podría inferir el valor de la variable Y conociendo el valor de X. Debido a estas ventajas, la correlación es una de las pruebas más usadas en el ámbito clínico, ya que además de medir la dirección y magnitud de la asociación de dos variables, es uno de los fundamentos de los modelos de predicción, como los modelos de regresión lineal, logística y riesgos proporcionales de Cox.

Palabras clave: Investigación clínica; Modelos de predicción; Correlación estadística

La cantidad de escritos de una profesión se correlaciona con su vitalidad y actividad, mientras que su calidad es un indicador aproximado de su estado intelectual

Sir Robert Hutchison (1871-1960)

Antecedentes

En artículos previos de la serie de Metodología de la investigación se abordó la manera de elegir una prueba estadística considerando el objetivo de la investigación, el tipo y distribución de las variables. Para determinar diferencia de medias en dos grupos no relacionados y variables cuantitativas de distribución normal se utiliza la prueba de t de Student, pero si lo que se quiere es predecir un desenlace dicotómico ajustando por distintas variables de confusión, la prueba estadística pertinente es el modelo de regresión logística múltiple. En este artículo abordaremos las pruebas estadísticas que permiten conocer la fuerza de asociación o relación entre dos variables cuantitativas u ordinales y cuyo resultado se expresa mediante el coeficiente de correlación. Si ambas variables se encuentran con distribución normal, calculamos la correlación de Pearson, si no se cumple este supuesto de normalidad o se trata de variables ordinales se debe calcular la correlación de Spearman[1,2]

Correlación de Pearson

El coeficiente de correlación de Pearson fue introducido por Galton en 1877 y desarrollado más adelante por Pearson. Es un indicador usado para describir cuantitativamente la fuerza y dirección de la relación entre dos variables cuantitativas de distribución normal y ayuda a determinar la tendencia de dos variables a ir juntas, a lo que también se denomina covarianza.

A continuación se muestra la fórmula para calcular el coeficiente de correlación de Pearson, la cual considera en el numerador la covarianza (suma de productos xy) y en el denominador, la raíz del producto de las sumas de cuadrados de ambas variables.

$$r_s = \frac{n\sum XY - (\sum x)(\sum Y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum Y^2 - (\sum Y)^2]}}$$

Ejemplo: para conocer la fuerza de asociación entre las variables del índice triglicéridos/glucosa y circunferencia abdominal, al ser variables cuantitativas de distribución normal la prueba estadística adecuada es la de correlación de Pearson.

Correlación de Spearman

La correlación de Spearman o también conocida como rho de Spearman es el análogo no paramétrico de la correlación de Pearson. Se utiliza para

variables cuantitativas de libre distribución o con datos ordinales. La correlación de Spearman se basa en la sustitución del valor original de cada variable por sus rangos, tal como se puede observar en su fórmula. Para calcularla se requiere que se ordenen los valores de cada sujeto para cada variable X, Y, además de que se asigne un rango. Si existe una correlación fuerte, los rangos deben ser consistentes: bajos rangos de X se correlacionarán con bajos rangos de Y. [3]

$$r_s = 1 - \frac{6 \sum d^2}{n\,(n^2 - 1)}$$

Donde:

$\sum d^2$ = sumatoria de la diferencia de rangos
n = número de pares (X, Y)

Ejemplo: si un investigador quiere determinar si existe correlación entre la saturación arterial de oxígeno y el estadio del pie diabético, la prueba estadística pertinente será la de correlación de Spearman.[4]

Interpretación y uso del coeficiente de correlación

El coeficiente de correlación se representa con una "r" y puede tomar valores que van entre −1 y + 1. Un resultado de 0 significa que no hay correlación, es decir, el comportamiento de una variable no se relaciona con el comportamiento de la otra variable. Una correlación perfecta implica un valor de −1 o + 1, lo cual indicaría que al conocer el valor de una variable sería posible determinarse el valor de la otra variable. Entre más cercano a 1 sea el coeficiente de correlación, mayor la fuerza de asociación (cuadro 1).

Además de analizar el coeficiente de correlación también debemos considerar el signo de este, que nos permite conocer la dirección de la correlación.[5] Un signo positivo implica que al aumentar la variable X también aumenta la variable Y, como ejemplo podemos encontrar que al aumentar la edad (X) aumenta la presión arterial (Y), mientras que un signo negativo implica que al aumentar la variable X disminuye la variable Y, tal como sucede con el índice tabáquico y el VEF_1: al aumentar el índice tabáquico (X) disminuye el VEF_1 (Y).

Las hipótesis que es posible plantearse mediante una correlación son las siguientes:

H_0: r = 0, no existe correlación.
H_a: r ≠ 0, existe correlación y está puede ser positiva o negativa.

Diagrama de dispersión

Para examinar visualmente una correlación se utiliza un diagrama de dispersión, el cual nos permite conocer el comportamiento de ambas variables. Cada uno de los puntos representa la intersección de un par de observaciones (X, Y). Con un suficiente número de datos podemos crear un diagrama de dispersión para observar la fuerza y dirección de la relación.

En la figura 1 puede apreciarse una correlación perfecta positiva; en este tipo de correlación es posible inferir un valor de Y si conocemos el valor de X, dado que se modifican constantemente. Por

Cuadro 1. Interpretación del coeficiente de correlación

0	Sin correlación
± 0.20	Correlación débil
± 0.50	Correlación moderada
± 0.80	Correlación buena
	Correlación perfecta

Parámetros solo de referencia, no deben ser considerados como estrictos puntos corte. Estos valores son afectados por el tamaño de muestra.

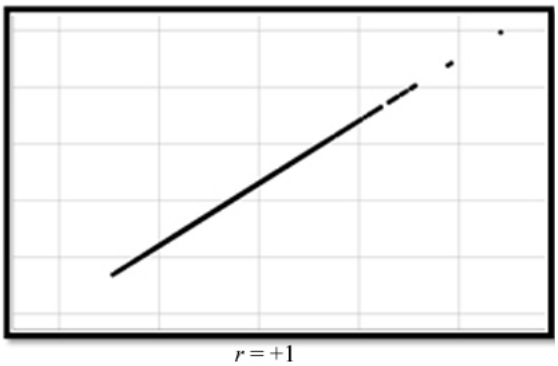

Figura 1. Correlación positiva.

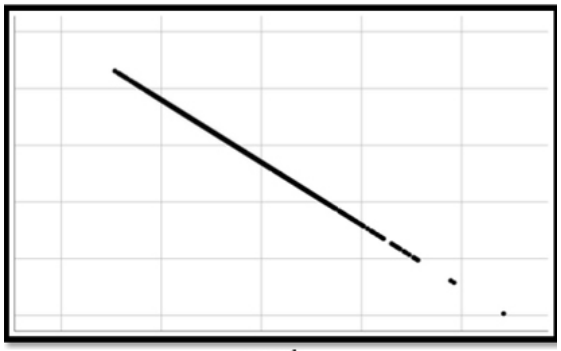

Figura 2. Correlación negativa.

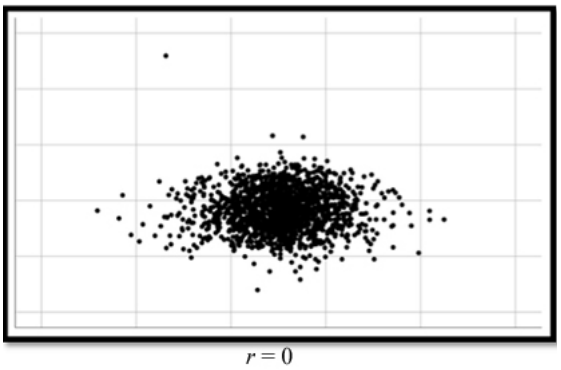

Figura 3. Diagrama de dispersión que muestra dos variables no correlacionadas.

ejemplo, si las variables edad y peso mostraran una correlación de + 1, con solo conocer el valor de edad podríamos determinar el peso de un niño. Sin embargo, es poco común encontrar este tipo de correlación, lo común es encontrar que las variables X y Y varían proporcionalmente y no siguen un patrón constante.

En la figura 2 se puede observar una correlación perfecta negativa. Es decir, al incrementar X disminuye el valor de Y, tal como podría ocurrir con las variables edad y masa muscular. Al aumentar la edad, disminuye la masa muscular.

En la figura 3 no es posible observar una relación lineal entre ambas variables. La variación observada entre una y otra es por efecto del azar. Esto ocurriría si quisiéramos mostrar la correlación entre consumo de chocolate al día y el coeficiente intelectual.

Significancia estadística del coeficiente de correlación

El coeficiente de correlación, así como otras pruebas estadísticas es dependiente del tamaño de muestra. Correlaciones de 0.20 pueden ser significativas con tamaños de muestras mayores, mientras que esta no será significativa si la muestra es pequeña, .[6]

Causalidad y correlación

Resulta necesario distinguir los conceptos de causalidad y correlación. La presencia de una correlación estadística entre dos variables no necesariamente implica causalidad, es necesario reflexionar acerca de algunas características que cuando se cumplen son sugerentes de una relación causal, como lo explicó sir Austin Bradford Hill varias décadas atrás. Por ejemplo, podríamos tener una correlación positiva y significativa entre las variables consumo de chocolate al día y el coeficiente intelectual, aún siendo significativa desde el punto de vista estadístico, no existe manera de explicar la plausibilidad biológica entre ambas variables.[7,8]

Coeficiente de determinación

Cuando exploramos una correlación lineal entre variables cuantitativas, parte de la variación de la variable Y puede ser debida a X. Sin embargo, en alguna proporción está variabilidad se deberá a otros factores o como efecto del azar. El coeficiente de determinación puede ser calculado para mostrarnos la proporción de variabilidad de la variable y que es atribuida a la relación lineal con X.

El coeficiente de determinación se obtiene elevando al cuadrado el valor del coeficiente de correlación (r^2). El coeficiente de determinación podrá tomar valores entre 0 y 1. Valores cercanos a 1 implican que una gran proporción de la variabilidad de y es explicada por X. Al llevar a cabo una correlación entre masa muscular y consumo máximo de O_2 en pacientes con insuficiencia cardiaca encontramos un coeficiente de correlación, $r = 0.7$ ($p < 0.0001$), el valor de R^2 será de 0.49, es decir, 49 % del consumo máximo de O^2 puede ser atribuido a la masa muscular.[9]

Ejemplo

Supongamos que el objetivo de un estudio es determinar si existe relación entre índice de masa corporal y porcentaje de grasa corporal por densito-

metría ósea. Nuestra hipótesis nula es que no existe relación y por lo tanto r = 0 y no existe correlación. La hipótesis alterna es que existe relación entre estas variables y que la asociación es positiva y por lo tanto r es distinto a 0.

A continuación, se muestran los pasos para la determinación de la correlación:

1. Determinar el tipo de distribución de las variables.
2. Al calcular las pruebas de normalidad encontramos que ambas variables tienen libre distribución, ya que la prueba de hipótesis de normalidad Kolmogorov-Smirnov muestra un valor de $p < 0.05$, por tanto, asumimos que nuestras variables tienen libre distribución (figura 4).
3. En virtud de que las variables de estudio tienen libre distribución, calcularemos la rho de Spearman.
4. En SPSS seleccionamos la opción de analizar y correlaciones bivariadas (figura 5).
5. Dado que nuestras variables presentan libre distribución, seleccionamos la opción de correlación de Spearman (figura 6).
6. Al analizar los resultados de la correlación, encontramos un coeficiente de correlación de

Pruebas de normalidad

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
IMC SUST. 1	.062	1688	.000	.960	1688	.000
porcent_grasa1	.045	1688	.000	.994	1688	.000

Figura 4. Resultado de la prueba de normalidad para índice de masa corporal (IMC) y porcentaje de grasa.

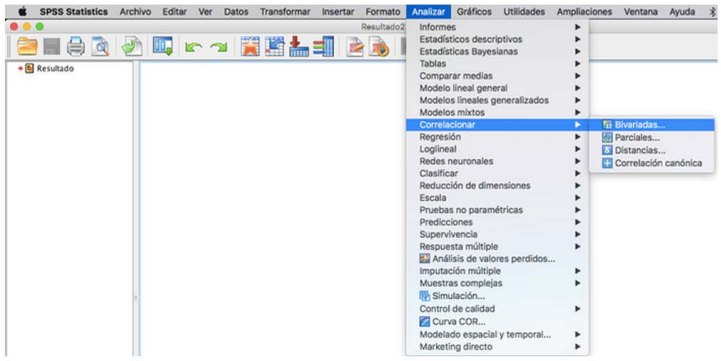

Figura 5. Pasos para llevar a cabo una correlación en SPSS.

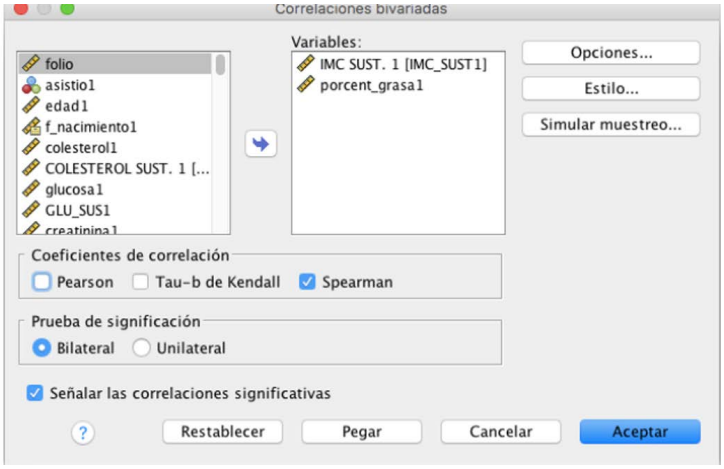

Figura 6. Pasos para llevar a cabo una correlación en SPSS.

0.69, con un valor de $p < 0.001$. Estos resultados nos muestran que la fuerza de asociación entre estas variables es de moderada a buena, se observa una correlación significativa, es decir que no es atribuible al azar. Al aumentar el índice de masa corporal se incrementa el porcentaje de grasa corporal (figura 7).

7. Para realizar el diagrama de dispersión, se elige la opción de cuadro de diálogos antiguos y elegirá diagrama de dispersión (figura 8).
8. La figura 9 nos muestra el diagrama de dispersión para las variables índice de masa corporal y porcentaje de grasa corporal, con una correlación positiva.

Correlaciones

			IMC SUST. 1	porcent_grasa1
Rho de Spearman	IMC SUST. 1	Coeficiente de correlación	1.000	.690**
		Sig. (bilateral)	.	.000
		N	1722	1688
	porcent_grasa1	Coeficiente de correlación	.690**	1.000
		Sig. (bilateral)	.000	.
		N	1688	1688

**. La correlación es significativa en el nivel 0,01 (bilateral).

Figura 7. Análisis de resultados de una correlación.

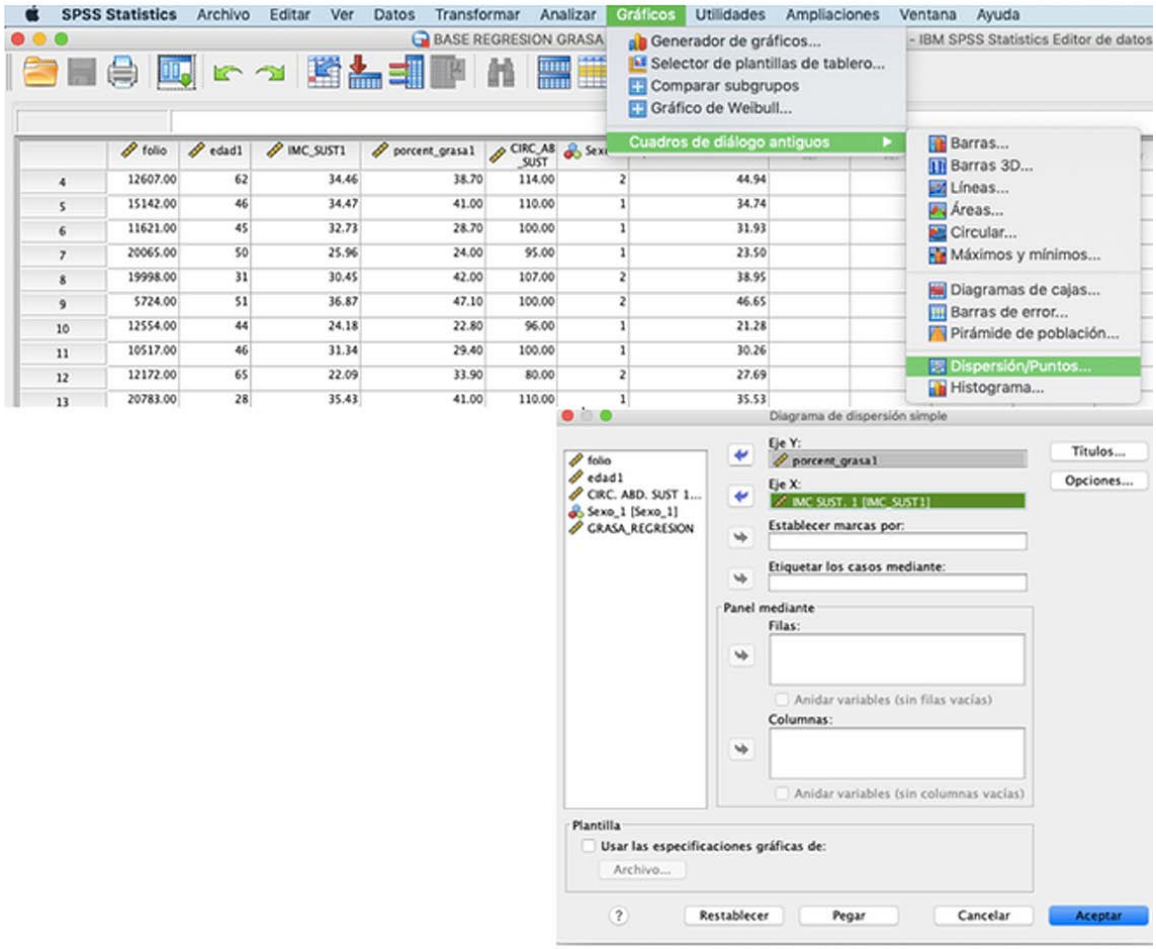

Figura 8. Pasos para llevar a cabo un diagrama de dispersión en SPSS.

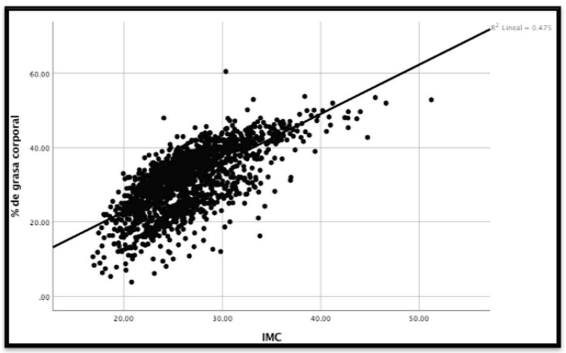

Figura 9. Diagrama de dispersión de la correlación entre índice de masa corporal y porcentaje de grasa corporal.

Otros usos

El empleo de los coeficientes de correlación es también útil para evaluar la relación entre las distintas variables independientes o covariables que planean ser introducidas en un modelo multivariado. De esa manera puede observarse si la correlación entre las variables independientes es alta (colinealidad) y evitar incluirlas juntas en un mismo modelo.

Conclusión

Las pruebas de correlación son de utilidad para identificar la tendencia de dos variables a ir juntas, lo que no necesariamente significa que cuando dos variables correlacionan entre ellas sea por una relación de causa y efecto. Hablar de causalidad es un tema mucho más complejo.

Agradecimientos

Este proyecto forma parte de la Iniciativa M (@iniciativaMp). Un especial agradecimiento al doctor Juan Talavera por ser nuestro mentor, incluso a la distancia.

Referencias

1. Talavera JO, Rivas-Ruiz R. Investigación clínica IV. Pertinencia de la prueba estadística. Rev Med Inst Mex Seguro Soc. 2011;49(4):401-405. Disponible en: https://www.medigraphic.com/pdfs/imss/im-2012/im121k.pdf
2. Flores-Ruiz E, Miranda-Novales MG, Villasís-Keever MÁ. El protocolo de investigación VI: cómo elegir la prueba estadística adecuada. Estadística inferencial. Rev Alerg Mex. 2017;64(3):364-370. DOI: 10.29262/ram.v64i3.304
3. Portney LG, Watkins MP. Foundations of clinical research. Applications to practice. EE. UU.: F.A. Davis Company; 2015.
4. Sánchez-López M, Roy-García I, Velázquez-López L, Navarro-Susano LG, Soriano-Pérez AM. Baja saturación de oxígeno como factor de riesgo para desarrollar pie diabético. Aten Fam. 2019;26(2):52-57. DOI: 10.22201/facmed.14058871p.2019.2.68826
5. Zou KH, Tuncali K, Silverman SG. Correlation and simple linear regression. Radiology. 2003;227(3):617-622. DOI: 10.1148/radiol.2273011499
6. Yazici S, Korkmaz U, Erkan M, Korkmaz N, Erdem-Baki A, Alçelik A, et al. The effect of breast-feeding duration on bone mineral density in postmenopausal Turkish women: a population-based study. Arch Med Sci. 2011;7(3):486-492. DOI: 10.5114/aoms.2011.23416
7. Talavera JO, Rivas-Ruiz R, Pérez-Rodríguez M, Roy-García IA, Palacios-Cruz L. De vuelta a la clínica: sin justificación no existe pregunta de investigación que valga. Gac Med Mex. 2019;155(2):168-175. DOI: 10.24875/GMM.19004942
8. Bradford-Hill A. The environment and disease: association or causation? Proc R Soc Med. 1965;58(5):295-300.
9. Cicoira M, Zanolla L, Franceschini L, Rossi A, Golia G, Zamboni M, et al. Skeletal muscle mass independently predicts peak oxygen consumption and ventilatory response during exercise in noncachectic patients with chronic heart failure. J Am Coll Cardiol. 2001;37(8):2080-2085. DOI: 10.1016/s0735-1097(01)01306-7