CSP CADERNOS DE SAÚDE PÚBLICA
REPORTS IN PUBLIC HEALTH

EDITORIAL
EDITORIAL

La importancia de la pregunta de investigación en el análisis de datos epidemiológicos

Cláudia Medina Coeli [1]
Marilia Sá Carvalho [2]
Luciana Dias de Lima [3]

doi: 10.1590/0102-311X00091921

[1] Instituto de Estudos em Saúde Coletiva, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil.
[2] Programa de Computação Científica, Fundação Oswaldo Cruz, Rio de Janeiro, Brasil.
[3] Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz, Rio de Janeiro, Brasil.

El modelado estadístico se usa frecuentemente para el análisis de datos epidemiológicos. Los modelos estadísticos son herramientas y se pueden emplear de forma diferente, dependiendo de si el objetivo de la investigación es una descripción, explicación causal o previsión [1]. Shmueli [1] plantea una discusión amplia sobre este tema, resaltando la importancia de adecuar la estrategia analítica a la pregunta de investigación.

El modelado descriptivo se usa para representar de forma económica la estructura de datos [1]. En Epidemiología, este modelado se utiliza cuando el interés es explorar la asociación entre varios factores de riesgo y un resultado. Se construyen modelos estadísticos con una selección de variables, basada en significación estadística y evaluación de ajuste del modelo [2]. Este tipo de estrategia todavía se adopta frecuentemente en artículos remitidos a CSP. Se usa incluso en temas para los que ya existen muchos artículos empleando el mismo planteamiento [3]. Otra limitación que se encuentra es la interpretación causal de las asociaciones observadas, inadecuada para este tipo de estudio.

En Epidemiología, el modelado explicativo se usa para comprobar hipótesis causales entre un factor de riesgo y un resultado. En el análisis, también se emplean modelos estadísticos, aunque la especificación del modelo está basada en el conocimiento *a priori* [4]. Se debe proponer un modelo teórico-operacional identificando, además de la exposición y el resultado, las variables de confusión y mediadoras. El modelo estadístico, entonces, se aplica a los datos para comprobar la hipótesis causal, teniendo como referencia el modelo teórico-operacional [1]. Algunos trabajos originales remitidos a CSP que prueban una hipótesis causal no orientan el análisis según un modelo teórico-operacional. Entre otros problemas, esto puede llevar a la inclusión indebida de covariables en el modelo estadístico, introduciendo un sesgo de selección [5]. En otros casos, se presentan y discuten resultados de medida de efecto, tanto para la variable de exposición, como para todas las covariables incluidas en el modelo estadístico. Esta estrategia es inadecuada, puesto que puede llevar a la interpretación incorrecta del efecto de las covariables (efecto total *versus* directo) [6].

Los trabajos originales que emplean un modelado predictivo son más raros en CSP. Como sucede en las Ciencias Sociales [7] y en la Psicología [8], en Epidemiología existe un mayor énfasis en la explicación causal que en la predicción. El modelado predictivo tiene como

objetivo predecir observaciones nuevas o futuras, empleándose tantos algoritmos de mineración de datos, como modelos estadísticos [1]. Incluso cuando se opta por los últimos, la estrategia analítica es diferente respecto a la que sería empleada cuando el objetivo es la explicación. Una cuestión central en el modelado predictivo es la validación cruzada, que permite evaluar la precisión del modelo en un conjunto de datos, diferente de aquel en el que fue probado [8]. En el modelado predictivo, no es necesario un modelo teórico-operacional muy elaborado. Por un lado, un modelo predictivo, aunque no represente adecuadamente la realidad, puede presentar un buen poder de predicción. Por otro, un modelo explicativo con un pequeño sesgo puede no presentar un buen poder predictivo [1]. Un problema hallado en los trabajos originales remitidos a CSP es el empleo del modelado descriptivo o explicativo con el objetivo de predicción. Otro problema observado es la utilización de toda la muestra, tanto para probar el modelo, como para evaluar la precisión de las predicciones.

La elección de la pregunta de investigación es la etapa esencial en la elaboración de cualquier trabajo original. Debe ser relevante, precisa y objetiva, orientando la estrategia analítica, así como la interpretación de los resultados alcanzados. En este sentido, en artículos que se apoyan en modelos estadísticos para el análisis de datos epidemiológicos es fundamental ser claros respecto a los objetivos de descripción, explicación o predicción de los fenómenos estudiados.

Colaboradores

Todas las autoras participaron en la redacción del texto y aprobación de la versión final.

Informaciones adicionales

ORCID: Cláudia Medina Coeli (0000-0003-1757-3940); Marilia Sá Carvalho (0000-0002-9566-0284); Luciana Dias de Lima (0000-0002-0640-8387).

1. Shmueli G. To explain or to predict? Stat Sci 2010; 25:289-310.
2. Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. 3rd Ed. Hoboken: Wiley; 2013.
3. Carvalho MS, Travassos C, Coeli CM. Mais do mesmo? Cad Saúde Pública 2013; 29:2141.
4. Hernán MA, Hernández-Díaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. Am J Epidemiol 2002; 155:176-84.
5. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004; 15:615-25.
6. Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. Am J Epidemiol 2013; 177:292-8.
7. Hofman JM, Sharma A, Watts DJ. Prediction and explanation in social systems. Science 2017; 355:486-8.
8. Yarkoni T, Westfall J. Choosing prediction over explanation in psychology: lessons from machine learning. Perspect Psychol Sci 2017; 12:1100-22.