CSP CADERNOS DE SAÚDE PÚBLICA
REPORTS IN PUBLIC HEALTH

EDITORIAL
EDITORIAL

A importância da pergunta de pesquisa na análise de dados 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.

A modelagem estatística é frequentemente usada para a análise de dados epidemiológicos. Modelos estatísticos são ferramentas e podem ser empregados de forma diferente, dependendo do objetivo da pesquisa ser descrição, explicação causal ou previsão [1]. Shmueli [1] faz uma discussão abrangente sobre esse tema, ressaltando a importância de se adequar a estratégia analítica à pergunta de pesquisa.

A modelagem descritiva é usada para representar de forma parcimoniosa a estrutura dos dados [1]. Em Epidemiologia, essa modelagem é utilizada quando o interesse é explorar a associação entre vários fatores de risco e um desfecho. São construídos modelos estatísticos com seleção de variáveis baseada em significância estatística e avaliação de ajuste do modelo [2]. Esse tipo de estratégia ainda é frequentemente adotada em artigos submetidos à CSP. É usada mesmo em tópicos para os quais já existem muitos artigos empregando a mesma abordagem [3]. Outra limitação encontrada é a interpretação causal das associações observadas, inadequada para esse tipo de estudo.

Em Epidemiologia, a modelagem explicativa é usada para testar hipóteses causais entre um fator de risco e um desfecho. Na análise, também são empregados modelos estatísticos, entretanto a especificação do modelo é baseada no conhecimento *a priori* [4]. Um modelo teórico-operacional deve ser proposto identificando, além da exposição e do desfecho, as variáveis de confusão e mediadoras. O modelo estatístico é, então, aplicado aos dados para testar a hipótese causal, tendo como referência o modelo teórico-operacional [1]. Alguns manuscritos submetidos à CSP que testam uma hipótese causal não orientam a análise segundo um modelo teórico-operacional. Entre outros problemas, isso pode levar à inclusão indevida de covariáveis no modelo estatístico, introduzindo viés de seleção [5]. Em outros casos, são apresentados e discutidos resultados de medida de efeito tanto para a variável de exposição como para todas as covariáveis incluídas no modelo estatístico. Essa estratégia é inadequada, uma vez que pode levar à interpretação incorreta do efeito das covariáveis (efeito total *vs.* direto) [6].

Manuscritos empregando a modelagem preditiva são mais raros em CSP. Como ocorre nas Ciências Sociais [7] e na Psicologia [8], na Epidemiologia há maior ênfase na explicação causal do que na predição. A modelagem preditiva objetiva predizer observações novas ou

futuras, sendo empregados tanto algoritmos de mineração de dados quanto modelos estatísticos [1]. Mesmo quando se opta pelos últimos, a estratégia analítica é diferente da que seria empregada quando o objetivo é a explicação. Uma questão central na modelagem preditiva é a validação cruzada, que permite avaliar a acurácia do modelo em um conjunto de dados diferente daquele em que foi treinado [8]. Na modelagem preditiva, não é necessário um modelo teórico-operacional muito elaborado. Por um lado, um modelo preditivo, mesmo que não represente adequadamente a realidade, pode apresentar um bom poder de predição. Por outro, um modelo explicativo com pequeno viés pode não apresentar um bom poder preditivo [1]. Um problema encontrado nos manuscritos submetidos à CSP é o emprego da modelagem descritiva ou explicativa com o objetivo de predição. Outro problema observado é a utilização de toda a amostra tanto para treinar o modelo como para avaliar acurácia das predições.

A escolha da pergunta de pesquisa é etapa essencial na elaboração de um manuscrito. Ela deve ser relevante, precisa e objetiva, orientando a estratégia analítica, assim como a interpretação dos resultados alcançados. Nesse sentido, em artigos que se apoiam em modelos estatísticos para análise de dados epidemiológicos é fundamental ter clareza quanto aos objetivos de descrever, explicar ou predizer os fenômenos estudados.

Colaboradores

Todas as autoras participaram da redação do texto e da aprovação da versão final.

Informações adicionais

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.