CSP CADERNOS DE SAÚDE PÚBLICA
REPORTS IN PUBLIC HEALTH

EDITORIAL
EDITORIAL

The importance of the research question in the analysis of epidemiological data

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

doi: 10.1590/0102-311X00091921

Statistical modeling is often used to analyze epidemiological data. Statistical models are tools that can be employed differently, depending on whether the research objective is the description, causal explanation, or prediction [1]. Shmueli [1] offers a comprehensive discussion of this topic, highlighting the importance of fitting the analytical strategy to the research question.

Descriptive modeling is used to represent the data's structure parsimoniously [1]. This modeling is used in Epidemiology when the interest is to explore the association between various risk factors and an outcome. Statistical models are built with the selection of variables based on statistical significance and evaluation of the model's fit [2]. This type of strategy is still frequently adopted in articles submitted to CSP. It is used even on topics for which there are already many articles employing the same approach [3]. Another common limitation is the causal interpretation of the observed associations, inadequate for this type of study.

Explanatory modeling is used in Epidemiology to test causal hypotheses between a risk factor and an outcome. The analysis also employs statistical models, but the model's specification is based on a priori knowledge [4]. A theoretical-operational model should be proposed identifying, in addition to the exposure and the outcome, the confounders, and mediators. The statistical model is then applied to the data to test the causal hypothesis, using the technical-operational model as the reference [1]. Some manuscripts submitted to CSP that test a causal hypothesis fail to orient the analysis according to a technical-operational model. Among other problems, this can lead to an undue inclusion of covariables in the statistical model, introducing a selection bias [5]. In other cases, the text presents and discusses results of effect measure both for the exposure variable and all the covariables included in the statistical model. This strategy is inadequate because it can lead to incorrect interpretation of the covariables' effect (total effect versus direct effect) [6].

Predictive modeling is much rarer in manuscripts submitted to CSP. As occurs in the Social Sciences [7] and Psychology [8], Epidemiology places greater emphasis on causal explanation than prediction. Predictive modeling aims to predict new or future observations, employing both data mining algorithms and statistical models [1]. Even when opting for the

[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.

latter, the analytical strategy differs from that employed when the objective is explaining. A central question in predictive modeling is cross-validation, which allows assessing the model's accuracy in a different dataset from that in which it was trained [8]. Predictive modeling does not require a highly elaborate theoretical-operational model [1]. On the one hand, a predictive model can have good predictive power even if it does not adequately represent reality. On the other hand, an explanatory model with a small bias may not have a good predictive power. A problem found in some manuscripts submitted to CSP is the use of descriptive or explanatory models for purposes of prediction. Another problem is the use of the entire sample to train the model and assess the predictions' accuracy.

Choosing the research question is an essential stage in a manuscript's elaboration. The question should be relevant, precise, and objective, orienting the analytical strategy and the results' interpretation. In articles that rely on statistical models for the analysis of epidemiological data, it is essential to clarify the objectives of describing, explaining, or predicting the phenomena in question.

Contributors

All the authors participated in writing the text and approval of the final version.

Additional informations

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.