The three main tasks of modern epidemiology: description, prediction and causal inference

José A. Calvache a-c, César Higgins Tejera d,e

a Editor in Chief, Colombian Journal of Anesthesiology. Bogotá, Colombia.
b Department of Anesthesiology, Universidad del Cauca. Popayán, Colombia.
c Department of Anesthesiology, Erasmus University Medical Center. Rotterdam, The Netherlands.
d Epidemiology Master’s Program, Universidad del Magdalena. Santa Marta, Colombia.
e Epidemiology Doctoral Program, University of Michigan. Michigan, United States.

Correspondence: Sociedad Colombiana de Anestesiología y Reanimación (S.C.A.R.E.), Cra 15a No. 120 – 74. Bogotá, Colombia.
E-mail: jcalvache@unicauca.edu.co

How to cite this article: Calvache JA, Higgins Tejera. The three main tasks of modern epidemiology: description, prediction and causal inference. Colombian Journal of Anesthesiology. 2023;51:e1088.

The advancement of science depends, at least partially, on editorial processes and the peer review. Despite multiple challenges and limitations, editorial and peer review processes continue to serve as quality filters for the improvement of scientific publications (1-3). As editors and reviewers, we work to improve the integrity and completeness of the report and discuss methodological and analytical issues that take part in the editorial process. At times, during the peer review process, we are able to identify inconsistencies between the research question, the study design, and the methodology of the study. These inconsistencies are critical elements that reviewers evaluate when assessing the viability of scientific publications (4).

The research question is one of the most important aspects of the scientific process. The research question must be clearly defined, because it informs the objectives of the study, an appropriate design, and a clear plan for analysis. As 2022 came to an end, the statistical editors of the British Medical Journal (BMJ) were looking forward to a quiet and peaceful Christmas holiday; with that in mind, they urged to pay attention to “twelve potential problems” they commonly identified as reviewers (5). At the top of their list was to have “absolute clarity of the research question”. The primary suggestion was to think carefully about the research question and be clear about the objectives of the study. This first step helps to characterize the study design (cross-sectional, longitudinal, etc.), and the measure of association (relative risk, odds ratio, prevalence risk) to be estimated (5). Missteps in this early phase of the research process can hardly be resolved by methodological adjustments and may lead to misinterpretations of the study results.

Briefly, there are three principal areas of modern epidemiology and data science: description, prediction, and causal inference (6). In biomedical research, the objectives that stem from the research question must be ascribed to one of these categories. In fact, these objectives are later translated into: 1) the selection of the study sample, which is characterized by the study population, place, and time; 2) the health outcome to be studied; 3) the measures of association to describe the event (incidence, prevalence, survival); and 4) the selection of a set of covariates that may be confounders of the relationship understudied (7). Given the current plethora of data, statistical softwares, and the advent of artificial intelligence, the aforementioned considerations are more relevant than ever.

To illustrate these concepts, we provide some examples recently published in the Colombian Journal of Anesthesiology.

The area of description employs data to provide a quantitative assessment, or a graphic summary, of certain characteristics of the world. Descriptive tasks include, for example, calculating a proportion — cumulative incidence or prevalence — of patients with postoperative nausea and vomit in a large hospital database or in a cohort study. Descriptive analyses range from basic summary calculations — mean and other measures of central tendency — to highly elaborated figures and sophisticated data synthesis techniques. For example, in a cross-sectional study, Bocanegra et al., (2022) provide a very clear description of the frequency of legal claims (closed cases) filed against anesthetists between 2013 and 2019 (8). Given the nature of the study design, these results cannot be generalized beyond the sample population. These limitations must be explicit within the study with the end to inform the
interpretations derived from it. In general, researchers must have a clear idea to what extent the objectives of their research are merely descriptive or seek other interests, because the objectives of the study must be reflected in the study design, the methodology, and the interpretation of the study results.

**Prediction** consists of using data to “map” certain characteristics of the world — inputs, or predictive variables — with other world characteristics — outputs or outcomes (6). Prediction usually begins with simple tasks, such as quantifying the association between midazolam premedication in children and the incidence of early postoperative delirium (9); and advances towards more complex tasks such as using multiple variables measured upon enrollment of patients undergoing cesarean section in order to “predict” which patients have a higher probability of developing postoperative nausea and vomiting (10). Predictive analyses range from simple calculations (e.g., incidence or risk difference) to more sophisticated modeling methods such as predictive and supervised learning algorithms (6). Questions related to “prediction or prognosis” are classified by the PROGnosis RESearch Strategy (PROGRESS) group (11) into four distinct types: 1) the ones that study the course of health-related conditions, or prognostic research; 2) the ones that study specific prognostic-related factors (biomarkers or others), or prognostic factors research; 3) the ones that study the development, validation and determination of the impact of statistical models on individuals’ disease risk and their future health outcomes, or prognostic models research; and 4) the ones that employ prognostic information for targeted individualized treatment decisions (11).

A common characteristic of predictive models is that the concept of “confounding bias” can become secondary because the primary focus of these models is not to establish “causal relationships” (11). However, advances in software development have enabled the integration of supervised learning models (like the Super Learner) as essential tools for estimating parameters of causal inference (12,13). The integration of these two areas holds significant promise for the advancement of epidemiological research in the 21st century.

**Causal inference** — defined by some authors as counterfactual prediction — uses data to predict certain features of the world, had the world been different; a journey back in time to change “something” in the past and observe what would have happened (6). The main aim of causal inference is to explain how the world works, and what would happen if we changed something in the world today. A widely known example of causal inference are randomized controlled clinical trials. In these studies, the random assignment of the intervention creates a counterfactual scenario where comparison groups are similar in terms of known and unknown characteristics that could influence the outcomes of the study. In a clinical trial, Casas-Arroyave et al., (2019) compared the use of a closed-loop system for the administration of total intravenous anesthesia versus the administration using a target-controlled infusion (TCI) (14). Many factors can influence the main outcome of this study, as is the case of the performance assessment of the system in terms of the depth of anesthesia, which is quantified using the bispectral index (BIS). However, those “factors” or confounding variables were controlled, in principle, by the methodological design of the experiment, and the randomized assignment of the treatment. This strategy allows us to recreate a “journey back in time”. In this journey, the same group of patients would have been subjected to the anesthetic procedure using TCI and assessed in terms of the health outcomes; later, the same group of subjects could “travel back in time” and be subjected to the closed-loop strategy. In the real world, we are only able to assess one of those potential outcomes; for this reason, causal inference problems tend to be seen as a missing data problem. In ideal conditions, the control group is used to assess what would have happened had the subjects in the study not been subjected to the study intervention, and this is what is meant by counterfactual prediction. This counterfactual reasoning represents the paradigm of epidemiological studies that employ causal inference, as the randomized trials (15,16).

The application of causal inference techniques in observational studies requires additional assumptions to those used in randomized trials (17-19). In some cases, the inherent limitations of observation studies (reverse causality, confounding bias) preclude the use of causal language when it comes to reporting and interpreting study results (20). Causality is a complex phenomenon that not only depends on the available information gathered in the data, it also requires external information, pre-existing knowledge, and the use of causal models that can be illustrated in the form of Directed Acyclic Graphs (DAGs). These graphs represent underlying premises, assumptions, theoretical concepts, and may provide guidance in the selection of confounding variables in regression models (21,22). Although some studies in the area of peripartum and intensive care (23-25) have approached causal inference using DAGs, the dissemination of these methods in such disciplines is still infrequent (26), and even more so in epidemiologic studies in Latin America. Therefore, this editorial is a call to study the counterfactual paradigm, and to implement causal inference methods in future epidemiological studies, with the objective to advance the national and Latin American scientific production toward the epidemiology of the 21st century.

It is worth highlighting that causal inference techniques are not only reserved for experimental trials; observational studies can also provide evidence regarding the ‘causal effects of interventions’ in cases in which a randomized trial is not feasible, ethical, or appropriate. However, making causal inferences from observational data is challenging due to confounding and selection biases, as well as other threats.
to the internal validity of observational studies. However, certain strategies such as Target Trial Emulation are being more accessible for solving causal questions in observational studies [27,28].

Finally, it is worth mentioning that the appropriate and clear selection of a research question leads to the correct interpretation of the study results. Clearly defining the aim of the study — including the meaning of variables such as “risk factors” or “predictive factors” — is essential for the correct and transparent interpretation of the study results, (29) and for avoiding causal misinterpretations or clinically irrelevant recommendations (5). It is not uncommon to find cases in which wrong causal interpretations are made on the basis of descriptive studies with obvious biases. We also believe that it is wrong not to call things by their name, and by what they seek to accomplish, if we set out to study causality and we use the appropriate methods to do so, we should not be afraid to use the word cause during the research process [30–31]. Straightforward questions and objectives help dispel the classical confusion between association and causality, a widely discussed conundrum in epidemiology, and a persistent topic in the scientific literature.

Funding

The authors declare not having received funding for the preparation of this article.

Conflicts of interest

The authors declare having no conflict of interest to disclose. JAC is the Editor-in-Chief for the Colombian Journal of Anesthesiology.

REFERENCES


