Statistical Modeling in Health Research: Purpose Drives Approach

Vielka González-Ferrer MD, Yainedy González-Ferrer DD, Marcos Ramírez-Marino MD

ABSTRACT

Statistical modeling is commonly used in both predictive and explanatory studies in health research. Its use in Cuba continues to grow, although it is sometimes employed inappropriately, which can lead to errors that imperil validity. This article attempts to shed light on faulty practices in statistical modeling by examining and discussing the main differences between explanatory and predictive models, with reference to the following: study objectives, theoretical considerations in model-building, aspects requiring assessment, variable and algorithm selection, analysis of confounders, treatment of multicollinearity, and reporting results.

KEYWORDS Prognosis, risk factor, protective factor, causality, statistical models, linear models, predictive models, explanatory models, logistic regression, Cuba

INTRODUCTION

Understanding health's complexities often requires research that examines multiple variables and their interrelationships. At the individual level, these include clinical and laboratory data and other attributes such as risk factors, and socioeconomic factors such as education. These studies commonly use statistical modeling as a resource to integrate multiple variables into a mathematical equation that portrays interrelationships among the variables. Two of the most common purposes of modeling are prediction and explanation.

Inappropriate model choice can lead to bias and misinterpretation. One of the most common errors in a predictive model is to use statistical variable selection algorithms to identify causes. In an enlightening paper on the use of instrumental variables for causal inference the authors say that "regardless of how immaculate the study design and how perfect the measurements, the unverifiable assumption of no unmeasured confounding of exposure effect is necessary for causal inference from observational data whether confounding adjustment is based on matching, stratification, regression, inverse probability weighting, or g-estimation."[1] Both linear (regression, analysis of variance and covariance)[2] and nonlinear (logistic and Poisson regression)[3] models are commonly used indistinctly to predict and to estimate causal effects without due attention to underlying assumptions.

A growing number of studies in Cuba are using statistical models for predicting events or identifying risk factors.[4–9] The common underlying drawback consists in using statistical criteria to identify relevant predictors and estimate measures of effect without a grounded theoretical analysis of their role in the models as true causes, confounders, effect modifiers or mediating variables. Although the formal structure of a predictive model may be similar to that of an explanatory model, to predict an occurrence is not the same as to explain its causes.[10] This article attempts to counter faulty practices in statistical modeling by examining and discussing the main differences between explanatory and predictive models.

Explanatory and predictive models in health: practical objectives When the dependent variable is binary (identifies whether an event occurs), the explanatory model includes a set of variables associated with probability of event occurrence (either as factors or markers of protection or risk). Whether a variable is a factor or a marker depends on the nature of its association with the dependent variable, which may or may not be causal. For example, smoking is a risk factor for lung cancer. But presence of yellow fingers—a common trait among inveterate smokers—is only a risk marker.

In an explanatory model for health, prioritizing variables helps inform and direct attention to the most important actions likely to mitigate or reduce risk. For example, if a study were to find that anemic pregnant women aged >35 years face 5 times greater risk of their newborn suffering congenital heart disease, that smoking triples risk, and that malnutrition doubles it, it would support issuance of guidelines to steer efforts toward reducing incidence of congenital heart disease by reducing the relative frequency of these factors in the population.

Unlike explanatory studies, predictive studies are used to inform physicians and patients about patients' health status and prognosis, enabling therapies and preventive actions to be fine-tuned to the individual. The term *prediction* is sometimes used incorrectly, especially when the temporal order implicit in the word "predictor" remains unverified. A so-called prediction may be a simple estimation. For example, a linear model could be used to estimate biparietal diameter as a function of gestational age, size of infarcted area as a function of concentration of an enzyme released during tissue lysis, or size of atheroma plaque in coronary arteries as a function of age and carotid Doppler results. None of these cases strictly presents a prediction; nor do they attempt to explain a process. In all these cases, the term *prediction* is being used incorrectly, instead of *estimation*.

Such tools have been used to predict treatment response in psychiatric illnesses,[11] susceptibility to preeclampsia,[12] and risk of hospital readmission,[13] among other things. Predictive models can also be applied in the social sciences, since they help identify subpopulations at risk, in order to focus actions on reducing or eliminating risk, managing resources based on scientific evidence and improving patient followup.

An explanatory model can be used for predictive purposes (depending on feasibility of practical application), but, as we will show in more detail below, a predictive model cannot always be used for explanatory purposes. An explanatory model has a theoretical cognitive underpinning not present in a predictive model, which is eminently practical, its purpose limited to prediction or estimation.

Model building: candidate variables and some remarks about statistical techniques Linear statistical modelling has become an important tool in predictive and explanatory studies because of its ease of interpretation. In the formal structure of a linear model,[1] each variable is multiplied by a coefficient, which, when standardized, directly measures the relative importance of the variable it accompanies.

Medical research often uses regression analysis techniques,[14] including binary logistic regression,[15] since outcomes are frequently expressed as dichotomous alternatives,[16,17] such as death (yes or no), risk of developing a disease (yes or no) and response to therapy (positive or negative).

A predictive model, as its name suggests, aims to make an accurate prediction with the greatest possible economy of resources. If a variable can be measured precisely, has good predictive capacity, and its inclusion does not affect the model's practical viability, then it usually will be included—regardless of whether it provides any information about causation. This does not mean that variables with good predictive capacity cannot play a causal role. Two good examples of predictive (but not explanatory) factors are skin coloration (in Apgar score) in neonatal prognosis and tumor markers of cancer prognosis or recurrence.[18]

Statistic modeling aiming at explanation—at estimating the causal effect of an explanatory variable on a response variable—must control for so-called *confounders*,[19–21] variables that are associated with both the explanatory and response variables but are not part of a causal pathway linking the two. Uncontrolled confounders tend to lead to biased estimates of causal effects. Controlling confounders is essential in explanatory models, but not for predictive models.

If, for example, the purpose is to study the relationship between age and dental caries, carbohydrate consumption could be a potential confounder. Because of its association with both variables (children consume carbohydrates more often than adults do; and frequent carbohydrate consumption increases risk of caries),[22] if this variable is not included in the model, results could mistakenly indicate that age is a protective factor for caries development (i.e., as age rises, risk of caries falls).

Criteria for model selection depend on the type of study. For a predictive study, a better model is one that will produce more reliable predictions. For a study aimed at investigating interrelationships among variables (correcting for the effect of others), a better model would be one that can obtain a more precise estimate of the coefficient of the variable of interest. The different goals of each type of study lead to distinctly different modeling strategies.[23] When regression techniques are used in an explanatory study, a variable that substantially modifies the value of the variable of interest's coefficient can be either a mediating variable or a confounder. Confounders should generally be included in the equation and mediating variables should not.[24] The relationship between the variable of interest and the probability the result will occur is observed to shift depending on whether that variable is taken into account. Including a mediating variable or excluding a confounder biases estimates of causal effects. In a predictive study, however, both can be excluded from the equation if they do not contribute to a more precise prediction or both can be included if they do.

Predictive models are built on the principle of parsimony, in the sense that if two models yield estimates or predictors of similar precision, the preferred model is the one with fewer predictors and fewer risk-modifying interactions. For prediction, it is advisable to limit use of interactions and include only those that are biologically plausible. For example, it is logical to think that people who consume many carbohydrates and also practice inadequate tooth-brushing would face greater risk of dental caries than those who exhibit only one of those two behaviors. Is it worth complicating the predictive model by including this interaction? Does variation in the model's performance justify its inclusion? To answer these questions, the model´s precision must be calculated with and without the interaction to determine which approach better predicts results for new subjects. Accordingly, algorithm use in selecting variables is fully justified for predictive, but not for explanatory studies.

Some authors consider that if a model performs well, the process of how it was obtained does not really matter.[25] Evaluation of a predictive model's performance consists of examining the accuracy of its predictions, usually by calibration and discrimination. Calibration measures the distance between predicted and observed results; for logistic regression models, this usually involves applying the Hosmer and Lemeshow test.[26] For binary models, it is important to determine the quality of discrimination between subjects who display the results described by the dependent variable and subjects who do not. A commonly used measure in binary models is the area under the receiver operating characteristic curve.[27] Assessment of a predictive model's performance may be overly optimistic if done with the same sample used to develop the model (training sample). A more realistic evaluation of model performance can be done by using a different subject sample (test sample). Goodness of fit for models used with an explanatory purpose is assessed by means of the percentage of variation explained (R^2).[28]

Collinearity, factor analysis, principal components analysis and reporting results When using regression techniques, researchers are concerned with the presence of two or more highly correlated variables. This phenomenon—called collinearity—can lead to large standard errors and biased estimates of model coefficients.

To detect collinearity, interrelationships among all explanatory variables are analyzed, and pairs of highly correlated variables are closely examined to decide whether one variable in the pair can be eliminated. In predictive studies, however, collinear variables can be helpful in reducing an estimate's standard error, so it is recommended that neither be eliminated. In the case of a high degree of collinearity among variables, dimension-reduction techniques are commonly used, which deliver a smaller number of mutually uncorrelated variables obtained as linear combinations of the original ones.[29,30]

Often there are many potential predictors and, in such cases, assessment of collinearity helps detect redundant information that can be eliminated (based on the principle of parsimony). For example, several anthropometric measures of pregnant women are collinear. If these are used to estimate newborn birth weight or to predict low birth weight, the marginal predictive capacity (reduction of the estimate´s standard error) of each variable when added to the ones already in the model should be assessed. If

Table 1: Differences in model application by type of study

Aspect	Type of study	
	Explanatory	**Predictive**
Practical goal	Estimate causal relationship between a dependent variable and a set of explanatory variables	Estimate risk to individual or population that a phenomenon will occur
Statistical techniques	Multivariate statistical classification techniques (often regression)	Multivariate statistical classification techniques or bivariate techniques
Essential considerations	Correctly estimate effect of causal factors or risk factors on results	Quantify performance and seek simplicity
Candidate variables	Explanatory variables	Predictors (explanatory or not)
Confounders	Essential to analyze	Not essential
Algorithms for selection of variables	Never justified	Justified
Treatment of multicollinearity	Change scale or eliminate some collinear variables	Change scale, eliminate some collinear variables or use other variables (factor or principal components analysis)
Reporting	Based on relative estimates of risk or change in response variable associated with changes in explanatory variables	Based on absolute estimates of risk or value of response variable

predictive capacity does not appreciably increase, the variable should not be included.

To address collinearity in explanatory studies, however, other solutions are often sought. These include transformations or changes in variable scale, standardization, or even elimination of certain collinear variables.[30] Use of factor analysis or principal components analysis is not appropriate because the specific purpose is to estimate the effect of the original variables on the response variable.

Finally, the nature of the model—the purpose it was created for—sets the course for study analysis and reporting. Predictive studies focus on absolute estimates of the probability that the result of interest will occur, or, if the dependent variable is continuous, on estimates of its magnitude. Causal associations and effects based on model coefficients have no direct relevance to building predictive models in practice. In contrast, explanatory studies usually aim to estimate causal effects represented by relative risks, interpreted as the quotient of the risks associated with presence or absence of the causal factor,[18] or in the case of continuous response variables, as the quotient of the model's coefficients as effect measures.

Table 1 displays differences in statistical models used for predictive versus explanatory studies.

CONCLUSIONS

Predictive and explanatory studies in health are particularly important due to the wide array of scenarios in which they can be applied. Depending on study type, multiple aspects change: purpose, analytic pathways for building and assessing models, and methods for interpreting results. This paper provides preliminary guidelines to help orient researchers who apply statistical models in health, contributors to the ever-growing body of Cuban—and international—scientific literature.

REFERENCES

1. Hernán MA, Robins JM. Instruments for causal inference: an epidemiologist´s dream? Epidemiol. 2006 Jul;17(4):360–72.
2. Rencher AC, Schaalje GB. Linear Models in Statistics. 2nd ed. New Jersey: John Wiley-Interscience; 2008 Jan 2. 688 p.
3. Lindsey JK. Nonlinear Models for Medical Statistics. Oxford (UK): Oxford University Press; 2001 Sep 20. 296 p.
4. Domínguez González EJ, Piña Prieto LR, Cisneros Domínguez CM, Romero García LI. Escala predictiva de mortalidad en la oclusión intestinal mecánica. Rev Cubana Cir. 2015 Apr–Jun;54(2):129–39. Spanish.
5. García Mederos Y, Zamora Matamoros L, Sagaró del Campo N. Análisis estadístico implicativo en la identificación de factores de riesgo en pacientes con cáncer de pulmón. MEDISAN. 2015 Aug;19(8):944–54. Spanish.
6. Bayarre H. Prevalencia y factores de riesgo de discapacidad en ancianos. Ciudad de La Habana y Las Tunas, 2000 [thesis] [Internet]. [Havana]: National School of Public Health (CU); 2003 [cited 2016 Apr 12]. 141 p. Available from: http://tesis.repo.sld.cu/70/1/Bayarre.pdf. Spanish.
7. Fuentes Díaz Z. Modelos multidimensionales pronósticos de mortalidad quirúrgica en intervenciones electivas no cardiacas [thesis] [Internet]. [Camagüey]: University of Medical Sciences of Camagüey; 2014 [cited 2016 Apr 12]. 131 p. Available from: http://tesis.repo.sld.cu/866/1/Zaily_Fuentes_D%C3%ADaz.pdf. Spanish.
8. Betancourt Cervantes JR. Nuevo índice predictivo para relaparotomías [thesis] [Internet]. [Havana]: Military Medicine Higher Institute of Havana; 2008 [cited 2016 Apr 24]. 87 p. Available from: http://tesis.repo.sld.cu/173/1/Betancourt_Julio.pdf. Spanish.
9. Jiménez Guerra SD. Modelo predictivo de neumonía y mortalidad en pacientes ventilados [thesis] [Internet]. [Matanzas (CU)]: Military Medicine Higher Intstitute of Matanzas (CU); 2008 [cited 2016 Jul 9]. 201 p. Available from: http://tesis.repo.sld.cu/204/1/Jiménez_Guerra.pdf. Spanish.
10. Shmueli G. To explain or to predict? Statistical Science. 2010;25(2):289–310.
11. Gupta M, Moily NS, Kaur H, Jajodia A, Jain S, Kukreti R. Identifying a predictive model for response to atypical antipsychotic monotherapy treatment in south Indian schizophrenia patients. Genomics [Internet]. 2013 Aug [cited 2015 Oct 2];102(2):131–5. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0888-7543(13)00018-9
12. Direkvand-Moghadam A, Khosravi A, Sayehmiri K. Predictive factors for preeclampsia in pregnant women: a receiver operation character approach. Arch Med Sci. 2013 Aug 30;9(4):684–9.
13. Billings J, Blunt I, Steventon A, Georghiou T, Lewis G, Bardsley M. Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (PARR-30). BMJ Open [Internet]. 2012 Aug 10 [cited 2015 Oct 2];2(4). pii: e001667. DOI: 10.1136/bmjopen-2012-001667. Available from: http://bmjopen.bmj.com/content/2/4/e001667.full
14. Lang H. Elements of regression analysis. Stockholm: KTH Mathematics; 2016 Jul. 58 p.
15. Berlanga-Silvente V, Vilà-Baños R. Cómo obtener un modelo de regresión logística binaria con SPSS. REIRE [Internet]. 2014 [cited 2017 May 15];7(2):105–18 . Available from: http://www.ub.edu/ice/reire.htm. Spanish.
16. Gispert Abreu EA. Morbilidad por caries dental y probabilidad de agravamiento en niños de 6 a 11 años [thesis] [Internet]. [Havana]: Higher Institute of Medical Sciences of Havana, School

of Dentistry; 2007. [cited 2016 Jul 9]. 187 p. Available from: http://tesis.repo.sld.cu/236/1/Gispert_Abreu.pdf. Spanish.

17. León Sánchez MA, Linares Guerra EM. La regresión logística binaria como instrumento para la predicción de deterioro inmunológico a partir de indicadores nutricionales en personas con VIH/SIDA. Rev Invest Operacional. 2014;35(1):35–48. Spanish.
18. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ. 2009 Feb 23;338:b375. DOI: 10.1136/bmj.b375.
19. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. 3rd ed. New York: Lippincott, Williams & Wilkins; 2009 Mar 1.
20. Fewell Z, Davey Smith G, Sterne JA. The impact of residual and unmeasured confounding in epidemiologic studies: A simulation Study. Am J Epidemiol. 2007 Sep 15;166(6):646–55.
21. Hernán MA. A definition of causal effect for epidemiological research. J Epidemiol Community Health. 2004 Apr;58(4):265–71.
22. González Sanz ÁM, González Nieto BA, González Nieto E. Salud dental: relación entre la caries dental y el consumo de alimentos. Nutr Hosp [Internet]. 2013 Jul [cited 2015 Oct 2];28(Suppl 4):64–71. Available from: http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0212-16112013001000008. Spanish.
23. Calderón Saldaña JP, Alzamora de los Godos Urcia L. Regresión logística aplicada a la epidemiología. Rev Salud Sexualidad Soc. 2009;1(4). Spanish.
24. Bacallao J. Mediating variable. In: Sarah Boslaugh, editor. Encyclopedia of Epidemiology. Vol. 2. New York: Sage; 2007:656–7.
25. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10(2):e1001381. DOI: 10.1371/journal.pmed.1001381.
26. Hosmer DW, Lemeshow S. Applied Logistic Regression. 2nd ed. New York: Wiley & Sons; 2000 Sep 15. 392 p.
27. Gonçalves L, Subtil A, Oliveira MR, de Zea Bermudez P. ROC curve estimation: an overview. REVSTAT [Internet]. 2014 Mar [cited 2017 May 17];12(1):1–20. Available from: https://www.ine.pt/revstat/pdf/rs140101.pdf
28. Nagelkerke NJD. A note on a general definition of the coefficient of determination. Biometrika.1991 Sep;78(3):691–2.
29. Velicer WF, Jackson DN. Component analysis versus common factor analysis-some further observation. Multivariate Behav Res. 1990 Jan1;25(1):97–114.
30. Hospital Universitario Ramón y Cajal [Internet]. Madrid: Hospital Universitario Ramón y Cajal; c2017. Material docente de la Unidad de Bioestadística Clínica. El problema de la colinealidad; 2010 [cited 2015 Mar 4]. Available from: http://www.hrc.es/bioest/Reglin_15.html. Spanish.

THE AUTHORS

Vielka González-Ferrer (Corresponding author: vielka@infomed.sld.cu), physician specializing in biostatistics, Ernesto Che Guevara Heart Center, Santa Clara, Cuba.

Yainedy González-Ferrer, dentist specializing in family dentistry, Celia Sánchez Manduley Dental Center, Santa Clara, Cuba.

Marcos Ramírez-Marino, physician with dual specialties in family medicine and obstetrics & gynecology, Mariana Grajales Maternity Hospital, Santa Clara, Cuba.

Submitted: January 18, 2017
Approved for publication: June 11, 2017
Disclosures: None

MEDICC Review A peer-reviewed, open-access journal

WRITE! REVIEW! READ!

Watch for our special issue on US–Cuban cooperation in April 2018!

Join the MEDICC Review community…

- Publish your research, reviews and evidence-based commentary
- Become a peer reviewer
- Join online and print readers from over 140 countries worldwide

Available in MEDLINE/PubMed, Thomson Reuters Web of Science, Elsevier's SCOPUS and EMBASE, CABI Global Health & Tropical Diseases Bulletin, Latindex and Redalyc

Online at
www.medicc.org/mediccreview

Manuscripts are accepted in Spanish or English.

Write editors@medicc.org for more information, or consult www.medicc.org/mediccreview for author instructions.

MEDICC Review publishes manuscripts—primarily from Cuba, Latin America and the Caribbean—for English-speaking readers concerned with innovations and strategies contributing to global health equity, cooperation and universal health.

The journal is a quarterly publication of Medical Education Cooperation with Cuba (MEDICC), Oakland, California, USA.