The Use of Statistical Learning in Pediatric Research

Byron C. Jaeger, PhD

In a recent study published in *The Journal of Pediatrics*, Scott et al[1] reported findings from an analysis that described the development and validation of a prediction model that detects septic shock in children by leveraging routine electronic health record data. The prediction model described in the authors' analysis predicts the probability of a pediatric patient with suspected sepsis experiencing septic shock during their medical encounter in urgent care, thereby increasing the chance of early detection and survival. The authors developed this prediction model by applying a statistical learning algorithm known as elastic net regression. The authors intentionally chose their approach to engage with electronic health records comprising numerous potential predictor variables for septic shock. Additionally, the authors applied core concepts from statistical learning to ensure that their prediction model would generalize to patients outside of their data.

See related article, *J Pediatr* 2020;217:145-51.e6

What Is a "Statistical Learning" Analysis?

Advances in data storage, data collection, and computational resources have allowed statistical learning (aka machine learning) to support clinical practice in certain settings.[2] Statistical learning analyses can be categorized as supervised or unsupervised learning.[3] Supervised learning (the framework applied by Scott et al) attempts to develop a function that predicts an outcome based on one or more predictor variables, whereas unsupervised learning attempts to uncover relationships and structure in data. The framework of supervised learning (ie, outcome modeled as a function of predictors) is similar to traditional statistical inference, for example linear and logistic regression in introductory statistics and biostatistics courses.[4] However, the primary aims and purposes of supervised learning usually differ from those of statistical inference. Briefly, statistical inference aims to determine whether an outcome (eg, survival) is associated with a predictor (eg, treatment). In contrast, supervised learning aims to optimize a model's external prediction accuracy (see Validation), with limited focus on individual associations between predictor variables and the outcome variable. Although some statistical learning algorithms create interpretable prediction functions (eg, elastic regression; see Learning Algorithms), others may be viewed as an uninterpretable "black box."[5] Others have previously argued that black box predictions may have unintended negative consequences in clinical practice (see Statistical Learning and Statistical Inference).[6] As such, methods to explain predictions from black box algorithms have become an important topic in statistical learning.[7]

Validation

The conceptual starting point of supervised learning is measuring the generalization error of a model. Generalization error is measured by assessing how well a model predicts outcomes for data that were not used to train it.[8] A limited analogy for model validation is teaching students (ie, models) with examples (ie, training data) and then performing a quiz (ie, testing data). For example, a teacher may tell students that 1 + 1 = 2, and that 1 + 2 = 3. Some students memorize the answers to their training examples, and others go further and learn the concept of addition. Both groups of students correctly answer their training examples, but only students who learned the underlying concept are able to successfully engage with testing examples such as 2 + 2 or 1 + 3. Similarly, model validation separates the models that have memorized (ie, overfitted) their training data from the models that have successfully identified a generalizable mapping from predictor variables to an outcome. In practice, it is recommended that models undergo both "internal" and "external" validation.

Internal Validation

Internal validation uses testing data from the same source as the training data. The simplest form of internal validation splits an initial dataset into 2 training and testing subsets. Although this strategy is more effective than testing a model using its own training data, there are a number of disadvantages.[9] A clear disadvantage is that holding a large sample out from the training data decreases the amount of data available to fit a prediction model. Another disadvantage is that split and test validation may depend heavily on the particular observations that were selected to be held out.

Resampling approaches have been developed to overcome the disadvantages of data splitting.[10] For example, Scott et al apply a 10-fold "cross-validation." This approach randomly partitions the training data into 10 nonoverlapping subsets (ie, folds) of roughly equal size. Next, 9 of the 10 folds are used to develop a model, which is then validated using data from the fold that was held-out. This process is replicated a total of 10 times, holding out a different fold in each replicate. A final estimate of model accuracy averages results from the 10 iterations. Notably, cross-validation does not validate a single model, but instead validates the algorithm used to

From the Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL.

The author declares no conflicts of interest.

https://doi.org/10.1016/j.jpeds.2020.04.005

develop a model.[11] Thus, cross-validation is often applied to identify an ideal modeling algorithm using a set of training data. The algorithm identified is then applied to the entire training dataset, and the final model is externally validated in downstream analyses.

External Validation

External validation involves testing data that are from a different (1) location, (2) population, or (3) time. Additionally, external data may be measured by a different set of instruments/technology or by a different research team.[9] Valuable insights can be drawn from using multiple types of validation data; for example, 1 validation set from a different time and another set from a different location. These analyses can diagnose specific ways that a model does or does not generalize. External data that can accurately test how a collection of models (some of which may leverage different training sets) would perform in "real-world" settings for a clinical prediction problem would be highly valued, but there are several challenges. It takes substantial effort and expertise to organize, label, de-identify, and safely share data from real-world clinical examinations, especially when follow-up is involved. Also, implicit or explicit biases in external data (eg, under-representation of minority groups) can lead to overly optimistic evaluation of a prediction model.

Learning Algorithms

Statistical learning algorithms comprise a diverse set of methods that leverage statistical principals and flexible optimization processes to form prediction functions. Well-known examples of statistical learning algorithms are elastic net regression, random forests, boosted models, and neural networks (ie, deep learning).[12-14] Each of these algorithms differs in technical detail, but has a consistent primary output: a prediction function that maps a set of inputs into a predicted output. This commonality allows learning algorithms to be objectively compared with one another (often in the context of cross-validation or external validation) by comparing the accuracy of their prediction function.

Statistical Learning and Statistical Inference

Inference and prediction are separate analytic procedures with different underlying assumptions and purposes. Clinicians should maintain awareness of these differences and exercise extreme caution when a statistical learning algorithm is used to conduct statistical inference. A common example is using stepwise variable selection (a statistical learning algorithm) to develop a model that is subsequently used to conduct inference regarding the variables selected. Other investigators have shown that standard errors and regression coefficients from this model are biased, and this substantially increases type I error rates.[9,15] Invalid inference occurs in this setting because of a failure to account for uncertainty in the variable selection procedure. For example, consider a scenario with millions of "noise" variables that have no association with an outcome of interest. By chance alone, some noise variables will be associated with the outcome in finite samples. Stepwise model selection select only these variables and compute standard errors that do not account for the millions of tests that were conducted to develop the model, thereby presenting *P* values that are highly biased. Selective inference procedures are a relatively new class of techniques that are meant to facilitate valid statistical inferences for this type of analysis.[16] For example, elastic net regression may provide valid statistical inference when the uncertainty of variable selection is taken into account and incorporated into standard errors of regression coefficients.[17]

Statistical Learning in Practice

The application of statistical learning algorithms in clinical settings presents challenges and opportunities related to interpretation and algorithmic bias. The guidelines for Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis should naturally be followed when the multivariable prediction model is based on a statistical learning algorithm, but some additional measures may need to be considered if the developed algorithm directly influences patient care.[18] Specifically, the algorithm's prediction function should meet criteria for interpretability and fairness. Regarding interpretability, the following information should be accessible for every patient's prediction: the contribution of each input to a model's prediction; the impact of changing input(s) on the model's prediction, holding other inputs fixed; and evidence quantifying the accuracy of the prediction model with external data.

Although elastic net regression naturally meets the first 2 criteria, other statistical learning algorithms do not. Specifically, the neural network can develop accurate prediction functions based on imaging data, but the resulting prediction function may contain thousands of parameters that are not easily interpreted. Strategies are being developed to bridge the gap between black box algorithms and interpretability.[19] A recent study by Poplin et al used neural networks to predict cardiovascular risk factors from photographs of the retina, and the authors showed that their neural network model could indicate which anatomic features (eg, the optic disc or blood vessels) were used to generate its predictions.[20] Statistical learning algorithms intended for use in clinical practice should also be vetted for algorithmic bias and practical usefulness. Algorithmic bias occurs when a model's training data contain systemic disparities between patient groups that impact patient outcomes. The model's authors should also consider the practical costs of medical errors (eg, implications of underdiagnosis or overdiagnosis) related to the task that their model supports and, if possible, estimate the impact of their model's clinical implementation in terms of those costs (eg, the number of adverse events prevented).

To summarize, statistical learning analyses should include details on the development and internal validation of a prediction function. Substantial effort should be taken to collect widely applicable external data for model testing. External validation should be performed to gauge the final prediction function's accuracy as well as diagnose biases in the model's prediction function. If statistical inference is performed, it is critical to account for uncertainty in the entire modeling process, and not just the final model. Detailed exposition on how one may interpret and explain individual results from the prediction function is critical if clinical implementation is a primary aim. ■

Reprint requests: Byron C. Jaeger, PhD, University of Alabama at Birmingham, Birmingham, AL 35209. E-mail: bcjaeger@uab.edu

References

1. Scott HF, Colborn KL, Sevick CJ, Bajaj L, Kissoon N, Deakyne Davies SJ, et al. Development and validation of a predictive model of the risk of pediatric septic shock using data known at the time of hospital arrival. J Pediatr 2020;217:145-51.e6.
2. Deo RC. Machine learning in medicine. Circulation 2015;132:1920-30.
3. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Springer; 2013. http://link.springer.com/content/pdf/10.1007/978-1-4614-7138-7.pdf. Accessed October 14, 2017.
4. Bzdok D, Altman N, Krzywinski M. Points of significance: statistics versus machine learning. Nat Methods 2018;1-7.
5. Castelvecchi D. Can we open the black box of AI? Nat News 2016;538:20.
6. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA 2017;318:517-8.
7. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv 2017;170208608.
8. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. Springer series in statistics. New York: Springer; 2001.
9. Harrell F. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. 2nd ed. New York: Springer International Publishing; 2015. www.springer.com/us/book/9783319194240. Accessed May 30, 2018.
10. Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001;54:774-81.
11. Kuhn M, Johnson K. Applied predictive modeling. New York: Springer; 2013.
12. Breiman L. Random forests. Mach Learn 2001;45:5-32.
13. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29:1189-232.
14. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 2017;19:1236-46.
15. Steyerberg EW, Eijkemans MJ, Habbema JDF. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol 1999;52:935-42.
16. Taylor J, Tibshirani RJ. Statistical learning and selective inference. Proc Natl Acad Sci U S A 2015;112:7629-34.
17. Lee JD, Sun DL, Sun Y, Taylor JE. Exact post-selection inference, with application to the lasso. Ann Stat 2016;44:907-27.
18. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Br J Surg 2015;102:148-58.
19. Molnar C. Interpretable machine learning. www.Lulu.com. Accessed 2019.
20. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng 2018;2:158.