Neural Networks, Logistic Regression, and Calibration: A Reply

To the Editor:—Steyerberg, Harrell, and Goodman[1] have commented about the model calibration, neural network design, and logistic regression models in our study, which used neural networks to predict perioperative cardiac risk.[2]

They maintain that the likelihood-ratio calculations were invalid unless the neural network scores had normal distributions for patients with and without cardiac events.[1] However, in our study we examined these distributions and they were indeed normal.

Obtaining likelihood ratios from neural networks was described as a superfluous step because neural networks can provide logistic probabilities.[1] However, many neural network designs do not provide such probabilities. Different transfer functions may produce neural network outputs greater than one or less than zero.

In our experience, the availability of different neural network transfer functions gives the investigator an important degree of versatility, with potential effects on model performance. With our method, the investigator does not need to rely on the NevProp software[3] of Dr. Goodman. The nature of the "more advanced functionality"[1] of NevProp[3] as compared with NeuralWare software[4] is unclear.

A concern about model calibration was also expressed. It was felt that the development of two separate models (one with six clinical predictors and a second with three dipyridamole thallium parameters) may have been inherently problematic.[1] However, in our experience with these data, we did not find better calibration through a single-step approach. Most important, the clinician may choose to defer dipyridamole thallium testing in patients whose six clinical predictors suggest low risk. The development of two separate models is preferable because it has greater clinical relevance.

The use of external validation was described as "quite common nowadays, but statistically inefficient compared to e.g. bootstrapping."[1] Again, clinical relevance is an issue. The ability to develop a model based on information from one hospital and implement it in evaluating data from another institution is a meaningful clinical issue that we addressed in our study. Careful attention to such clinical issues shaped many of the other design choices in the original model development and validation provided by L'Italien et al.[5]

Steyerberg, Harrell, and Goodman provide further comments on the development of logistic regression models, proposing alternative methods and suggesting that in comparison neural networks might not have seemed so impressive.[1] This type of response to neural network studies is quite common. The reader feels that if only a more advanced statistical method had been employed, then the performance difference might have been eliminated. This approach misses the point of our work.

Our goal was to examine a previously implemented, successful, published statistical approach[4] and determine whether a neural network model could offer any differences in model performance. We feel that relevant differences were indeed shown. The neural network therefore offers one option for model refinement. This does not mean that there are no other statistical options with logistic regression. Nor can we exclude other neural network modeling options that were not evaluated in the study.

This research should not be envisioned as a contest between the best possible neural network and the best possible statistical model, but rather as a study of how an investigator might use neural networks to refine model performance. In this respect, a neural network approach may not be inherently "superior," but with existing software it may indeed be practical and useful. Neural networks have attracted substantial interest because they readily provide the PC user with powerful and advanced models. Neural networks merit further development and evaluation in the field of medical decision making.

PABLO LAPUERTA, MD
GILBERT L'ITALIEN, PhD
Department of Outcomes Research
Pharmaceutical Research Institute
Bristol–Myers Squibb
Princeton, New Jersey

ROBERT GIUGLIANO, MD, SM
Cardiovascular Division
Brigham and Women's Hospital
Harvard Medical School
Boston, Massachusetts

KIM A. EAGLE, MD
Cardiology Division
University of Michigan Medical Center
Ann Arbor, Michigan

References

1. Steyerberg EW, Harrell FE, Goodman PH. Neural networks, logistic regression and calibration (letter). Med Decis Making. 1998;18:349–50.
2. Lapuerta P, L'Italien GJ, Paul S, et al. Neural network assessment of perioperative cardiac risk in vascular surgery patients. Med Decis Making. 1998;18:70–5.

3. Goodman PH, NevProp3: Artificial Neural Network Software for Statistical Prediction. Reno, NV: University of Nevada, 1996. Available at ⟨http://www.scs.unr.edu/nevprop/⟩.
4. Neural Computing: A Technical Handbook for Professioin II/Plus and NeuralWorks Explorer. Pittsburg, PA: NeuralWare, Inc., 1993.
5. L'Italien GJ, Paul SD, Hedle RC, et al. Development and validation of a Bayesian model for perioperative cardiac risk assessment in a cohort of 1,081 vascular surgical candidates. J Am Coll Cardiol. 1996;27:779–86.

Neural Networks, Logistic Regression, and Calibration: A Rejoinder

To the Editor:—In closing this discussion about calibration of neural networks (NNs) and logistic regression, it is not necessary to reiterate all criticisms mentioned before. The reader can contemplate some of the minor issues, such as the pros and cons of using specific neural network software.[1] Rather, we would like to expand on the more general issues on prognostic modeling with NNs or logistic regression techniques.

Lapuerta et al. reply that their research should not be envisioned as a contest between the best possible neural network and logistic regression model.[2] In their study, a comparison was made between a poorly validating neural network and a very poorly validating logistic regression model; indeed, these models were far from the best possible, for reasons indicated before. It makes little sense to compare methodologic aspects of two techniques when at least one is applied in a simplistic way. One should not ignore the advances in regression modeling that have been made over the past 10 years.[4]

Regarding model validation, our remarks on bootstrapping should not be misunderstood as not clinically relevant. Statistical issues related to internal validity should be distinguished from clinical issues relating to external validity. Internal validity refers to the performance of the model (e.g., calibration, ROC area) in the underlying patient population from which the sample used to construct the model was drawn. Bootstrapping is currently the most efficient way to assess internal validity. A bootstrap estimate of internal validity will be less favorable than the apparent validity estimated directly on the sample. The latter estimate is over-optimistic, especially when the data set is small.

External validity, or generalizability, refers to validity in another population, for example, in patients seen more recently (time difference), or at different centers (place difference). Several mechanisms (referral patients, registration of predictors, treatment differences, etc.) may lead to differences in prognostic relationships, causing external validity to be most likely less than internal validity. Internal validity will imply external validity only when differences in time and place are negligible. When time or place effects are present, they can be incorporated in the model so that valid predictions can be obtained for another population.[5]

In the original publication of L'Italien et al. it was argued that comparability may be assumed between training (two hospitals) and validation samples (three other hospitals).[6] This implies that the patients in the five hospitals were regarded as originating from a common underlying population. The evaluation thus concerned essentially internal validation, and bootstrapping of a model based on the full data set would have been more efficient.

If one is interested in assessing external validity, internal validity should first be assessed in a training sample (e.g., by bootstrapping). Subsequently one might compare the estimated internal validity with the estimate in a validation sample. Large sample sizes will be required to determine differences between internal and external validity with some certainty.

In conclusion, assessment of internal and external validity should not be mixed up in one step, and the bootstrap has a clear place in model validation.[4] When modern methods such as NNs are applied, an up-to-date approach to regression analysis should be used in any comparisons. Any advantages of NNs may then appear to be restricted to patient populations with highly nonlinear prognostic relationships and important higher-order interactions that are not easily captured in regression models.[1,7]

Ewout W. Steyerberg, PhD
Department of Public Health
Erasmus University
Rotterdam, The Netherlands

Frank E. Harrell, Jr., PhD
Department of Health Evaluation Sciences
University of Virginia
Charlottesville, Virginia

Philip H. Goodman, MD, MS
Department of Internal Medicine
University of Nevada
Reno, Nevada

References

1. Goodman PH. NevProp3: Artificial Neural Network Software for Statistical Prediction. Reno, NV: University of Nevada; 1996. Available at: ⟨http://www.scs.unr.edu/nevprop/⟩.
2. Lapuerta P, L'Italien G, Giugliano R, Eagle K. Neural networks