EDITORIAL

Statistical Prediction Models, Artificial Neural Networks, and the Sophism "I Am a Patient, Not a Statistic"

WHEN I WAS diagnosed with lymphoma 11 years ago, I was eager to learn my prognosis. As a graduate student, I had excellent electronic access to the medical literature and was quickly able to review numerous statistical analyses of patients with my disease. A consistency of these analyses was that each provided a stage-specific prognosis, without adjustment for other prognostic factors. For example, I had heard I had a favorable histologic subtype, but I did not know how much weight, if any, it should be given. And although I was stage IV, I was told I was more of a "bad" stage II than a typical stage IV, so I figured my extent of disease might be less than average for my stage. Thus I wondered if I should interpolate between the stage II and IV prognoses, and try to mentally average these Kaplan-Meier curves to arrive at a prediction tailored to my disease and me. I really wanted a predicted probability of survival and didn't specifically care what the prognostic factors were, what my relative risk might be, or in what risk group I belonged. I had no desire to understand the prognostic model, but I demanded the most accurate prediction currently available. I had great difficulty trying to reconcile the mindset of the business school, where I was studying, with apparently that of the medical school, where I was now living. In the business school, the mentality is that if you can predict future outcomes more accurately than the next person, you win, end of story. I personally felt (and still do feel) that this attention to predictive accuracy has a place in medicine.

Djavan et al[1] clearly appreciate predictive accuracy. In the current issue of the *Journal of Clinical Oncology*, they compare an artificial neural network (ANN) prediction model with traditional logistic regression in their ability to predict, on the basis of several diagnostic tests, whether a man has prostate cancer. Artificial neural networks get their name from the perception that they imitate the natural neural networks in our heads. Regardless of whether this is true, it is not clear that we would want to replicate our natural neural network on the computer for the purpose of prediction. Most studies have shown that human experts are generally of inferior accuracy when compared with predictions made by typical statistical models.[2] In any case, one can think of an ANN as a very complicated regression equation (eg, prediction = coefficient 1 × predictor variable 1 + coefficient 2 × predictor variable 2. . .). What makes the ANN particularly complicated is the fact that there are intermediate predictions (intermediate prediction 1 = coefficient 1 × predictor variable 1. . .), which themselves make the final prediction (prediction = intermediate coefficient 1 × intermediate prediction 1. . .). In other words, the original predictor variables (such as laboratory tests) are combined to form an intermediate prediction, and then the intermediate predictions are combined to calculate the final predicted probability of whether the patient has the disease of interest. These intermediate predictions, which form the hidden layer of the ANN, make the ANN more flexible than logistic regression.

Djavan et al[1] took a large data set and split it into three parts. With the first part (training, 50% of the original data set), an ANN prediction model was developed. The second part of the data (testing, 25%) was used to help guide the fitting of the ANN, helping it to achieve its maximum accuracy. The third part (validation, 25%) was used to evaluate the performance of the ANN. The competing method for prediction, a logistic regression model, was built using two thirds of the data and tested on the remaining third. From a study design perspective, it might have been more interesting to give both techniques equal amounts of data from which to arrive at a final model, instead of giving three quarters to the ANN and two thirds to logistic regression. This involved process of splitting the data and comparing the techniques, like that performed by Djavan et al, is necessary because it is extraordinarily difficult to predict from characteristics of the data whether a machine-learning approach like the ANN will outperform an ordinary statistical method.[3] In a comprehensive review of ANN and statistical method comparisons in oncology diagnosis and prognosis, Schwarzer et al[4] made a convincing argument that there is no evidence thus far that ANNs represent real progress. They raised numerous concerns about the studies performed to date, challenging whether many were really fair contests and judged in an unbiased fashion. Clearly, a huge practical advantage of logistic regression over an ANN, in the setting of equivalent accuracy, is that the logistic regression prediction model is easy to represent on a piece of paper[5] without necessarily having to resort to software for implementation.

There are reasons why an ANN should outperform logistic regression. Figure 1 illustrates the main ANN advantages. This figure is a hypothetical plot (a sample space) of two diagnostic tests, percentage of free prostate-specific antigen (PSA) versus total PSA. In this plot, P represents a patient with prostate cancer and N represents a biopsy-negative patient. A classification technique attempts to separate the P's from the N's by dividing the sample

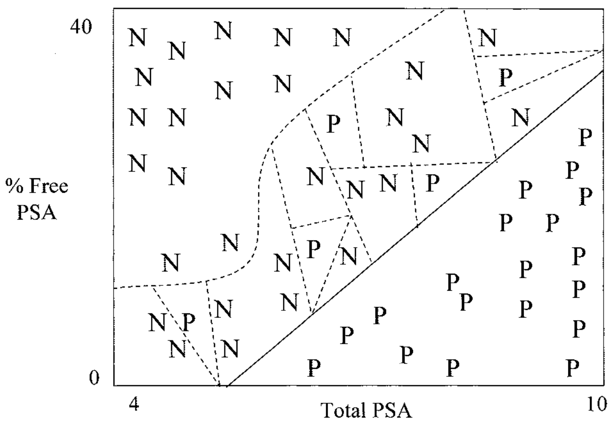

Fig 1. **In this hypothetical example, both logistic regression and the ANN have drawn the solid line, but the ANN has also drawn the dashed lines. Assuming the N's and P's are randomly distributed in the diagonal region where they overlap, the extra lines reduce predictive accuracy for future patients and represent overfit.**

space into regions associated with either a positive or negative biopsy. The key is to partition the sample space such that the predictive accuracy is maximized when attempting to predict the outcomes of new patients, not necessarily those in the sample space. Logistic regression in its ordinary formulation (without interaction terms, polynomials, and so on), can draw only a single straight line in an attempt to separate the N's from the P's. An ANN can draw an unlimited number of linear or nonlinear boundaries in its attempt to create regions of positivity or negativity.[6] This flexibility is the primary strength of the neural network but also the potential weakness. Drawing too many lines in the sample space produces overfit, which in turn reduces predictive accuracy.[7] For example, suppose that a region of the sample space contains positive and negative cases randomly distributed, with 75% of the cases being negative (Fig 1). Extreme overfit would form homogeneous regions with only positive or negative cases, resulting in approximately 75% of the sample space allocated to negative cases for this example. Thus, when used to predict future cases, the sample space will predict negative approximately 75% of the time. Assuming future cases are, in fact, negative approximately 75% of the time, the overfitted sample space is expected to achieve $0.75 \times 0.75 + 0.25 \times 0.25 = 0.625$ accuracy. In words, the sample space will predict negative 75% of the time, and 75% of the cases will be negative; the sample space will predict positive 25% of the time, and 25% of the cases will be positive. The reason overfit results in inferior predictive accuracy becomes apparent when considering how accurate prediction would have been if no lines were drawn at all. With no lines, the sample space would always predict the dominant class (negative), and thus achieve 75% accuracy ($1 \times 0.75 + 0 \times 0.25 = 0.75$), which is better than the 62.5% accuracy achieved from overfit. If multiple or nonlinear partitions are necessary, and the ANN draws these and only these, it should predict more accurately than logistic regression when applied to new patients, because logistic regression cannot draw these partitions. If the ANN does not detect the need for the multiple or nonlinear partitions, it will likely tie logistic regression because it will likely partition the sample space in the same manner as logistic regression. If the ANN draws too many partitions, it likely will lose the competition because of the overfit problem described above. In Fig 1, drawing the dashed lines would likely diminish predictive accuracy, because they seem to represent overfit. The area of overlap would have best been left fully allocated to the dominant class: negative.

The study by Djavan et al[1] also highlights how difficult it is to compare models with respect to predictive accuracy. Many would agree with Djavan et al that area under the receiver operating characteristic curve (ROC) is the metric of choice for judging discrimination when predicting binary outcome data, such as presence or absence of prostate cancer. By that metric, neither of the two ANNs in this study (one for PSA between 2.5 and 4 ng/mL, the other for PSA between 4 and 10 ng/mL) is statistically significantly better than its logistic regression counterpart. That is a frustrating conclusion when the neural network ROC curve seems to dominate much of the logistic regression ROC curve (Figs 3 and 4 in Djavan et al). Although it is tempting to compare sensitivity or specificity at a particular cut point, Harrell[8] cautions against this for several reasons related to the dependency of the conclusions on the specific cut point selected, instead deferring to the full area under the ROC curve as the preferred metric. Rather than selecting a single validation subset of the data, Schwarzer et al[2] encourage the use of cross-validation to increase the efficiency of the analysis. For example, 10-fold cross-validation splits the data set into 10 equal parts. Each 10th is used for testing a model derived from the remaining nine parts. This process of developing a prediction model with 90% of the data and testing with the remaining 10% is repeated for a total of 10 accuracy observations. Perhaps cross-validation would have provided the statistical power increase necessary to declare the ANN as more accurate than logistic regression, or vice versa.

The area under the ROC curve provides a useful measure of discrimination, how well a prediction model can rank patients. However, it does not provide much insight into calibration, which refers to the correspondence between predicted and actual probabilities. Calibration curves, which are plots of actual against predicted probabilities, are very useful for visually determining accuracy. One would generally like to see such a curve before providing predicted probabilities to a patient.

Predictive models applied to the individual patient are sometimes controversial. The chief complaint against them is that an individual is a patient, not a statistic. The argument is that statistics apply to groups of patients and not to individual patients. For example, an individual patient either has prostate cancer or not, and a probability of having prostate cancer has no useful interpretation for him. This seems like a weak argument, and some extreme examples show why. Suppose you are forced to play Russian roulette and have the choice between two six-shooters; one has one bullet, and the other has five. Would you be indifferent in your choice? You will either survive or not, but I have a feeling you will choose the pistol with a single bullet. Is the weather forecast helpful to you when deciding whether to carry your umbrella? It will either rain or not today. When comparing two surgeons in their ability to perform a procedure that you need, would it be helpful to your decision making to learn that, when operating on patients who appear to be nearly identical to you, one surgeon has an operative mortality rate of one in 1,000, whereas the other has a rate of 30%? Each of these examples illustrates that imperfect prediction, despite being imperfect, can be valuable for decision-making purposes. True, a binary outcome will either occur or not, but the hypothetical rate at which it would occur if the experiment could be repeated is of value to the decision maker.

If imperfect predictions are valuable for decision making, more accurate predictions should be preferred to less accurate predictions. With prostate cancer detection, decision rules are implicitly in place (eg, many use PSA $\geq$ 4.0 ng/mL or abnormal digital rectal examination as reason to biopsy). A rule that predicts more accurately could detect more cancers, produce fewer negative biopsies, or both. Thus we should strive to produce increasingly accurate prediction models by increasing sample sizes from which to build prediction models, adding informative markers, and applying more sophisticated modeling approaches. Improving our predictive accuracy is critical not only in prostate cancer, but in all cancers. Cancer patients deserve and expect improved prediction.

Michael Kattan
Memorial Sloan-Kettering Cancer Center
New York, NY

REFERENCES

1. Djavan B, Partin AW, Remzi M, et al: A novel artificial neural network for early detection of prostate cancer. J Clin Oncol 20:921-929, 2002

2. Dawes RM, Faust D, Meehl PE: Clinical versus actuarial judgment. Science 243:1668-1674, 1989

3. Kattan MW, Cooper RB: The predictive accuracy of computer-based classification decision techniques: A review and research directions. Omega Int J Mgmt Sci 26:467-482, 1998

4. Schwarzer G, Vach W, Schumacher M: On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. Stat Med 19:541-561, 2000

5. Eastham JA, May R, Robertson JL, et al: Development of a nomogram that predicts the probability of a positive prostate biopsy in men with an abnormal digital rectal examination and a prostate-specific antigen between 0 and 4 ng/mL. Urology 54:709-713, 1999

6. Hornik K, Stinchcombe M, White H: Multilayer feedforward networks are universal approximators. Neural Networks 2:359-366, 1989

7. Kattan MW, Cooper RB: A simulation of factors affecting machine learning techniques: An examination of partitioning and class proportions. Omega Int J Mgmt Sci 28:501-512, 2000

8. Harrell FE Jr: Regression Modeling Strategies With Applications to Linear Models: Logistic Regression, and Survival Analysis. New York, NY, Springer-Verlag, 2001