Why Models Predicting Bacteremia in General Medical Patients Do Not Work

In this issue of the Journal of General Internal Medicine, Yehezkelli and his team describe the prospective validation of two clinical rules for the early prediction of bacteremia.[1] They conclude that both models showed marked deterioration in performance over the initial reports and emphasize caution in transporting the results to other populations. Understanding the significance of this article requires a broader view of clinical prediction rules in general and previous work in predicting bacteremia in particular. Several models have been developed for predicting bacteremia, but their utility remains debatable. The authors of this current study chose two of the most popular models.

The first model was derived by Bates from a prospective cohort of 1,318 patients. His team sampled all patients with blood cultures drawn at the Brigham and Women's Hospital during a several month period from 1988-89.[2] In this original population, 1% of the low-risk group had bacteremia compared to 16% of the high-risk group. Predictive factors included temperature, type of disease, presence of chills, history of intravenous drug use, acute abdomen by physical exam, and major comorbidity.

In contrast, the second model was developed by Leibovici using 244 patients consecutively hospitalized because of a febrile illness.[3] Using this model, low-risk patients had a 5% incidence of bacteremia compared to an incidence of 83% for high-risk patients. Predictive factors included serum albumin, presence of chills, functional status by a modified Karnofsky scale, admission diagnosis of urinary tract infection, and renal failure.

Other models predicting the occurrence of bacteremia have been developed.[4–8] Three authors developed models for general medical patients.[9–11] In general, patients determined to be at low risk had an incidence of bacteremia ranging from 2% to 15%. High-risk patients had bacteremia rates ranging from 16% to 33%. The model by Moses and his team proved to be inconsistent from season to season even at the same hospital.[10] Seasonal variation in the proportion of Staphylococcal and Pseudomonas organisms contributed to this instability. Pfitzenmeyer and his team found that the subjective judgment of physicians had operating characteristics very similar to their prediction rule.[11]

The literature contains many applications of these models.[12] Proposed uses include assisting the physician in making clinical decisions involving the need to (1) obtain blood cultures, (2) give empiric antibiotic treatment, and (3) aggressively monitor for clinical deterioration. Others have focused on using the models to interpret positive blood cultures.[13]

These models have even been proposed as guidelines for quality-improvement and cost-containment efforts. For example, blood cultures and empiric antibiotics may be withheld from low-risk patients defined by these models. Schwenzer noted that among all patients who had blood cultures collected, the cost of a blood culture with a clinical impact was $4,622 per patient,[5] because of the low yield in the intensive care setting. Reducing the number of blood cultures assumes that a clinician feels comfortable not testing or treating a patient otherwise thought to be at risk for bacteremia. Although no formal decision analysis has established a threshold for withholding blood cultures or empiric antibiotics in the setting of suspected bacteremia, we believe that the threshold should be lower than the percentage of bacteremia in the low-risk patients from most of these studies.

Understanding the inadequacies of these models requires a more in-depth discussion of clinical prediction rules in general.[14] There are three important methodological questions about discrimination, transportability, and calibration. The first question is, "Does the model discriminate between disease and non-disease?" The area under the receiver operating characteristic (ROC) curve gives the best measure of discrimination.[15] To illustrate this point, consider two populations. The first population consists of all bacteremic patients admitted to the medical ward of a hospital during one year. The other population consists of similar patients without bacteremia. Now, randomly select a series of pairs with one member from each population and apply the prediction rule to determine who is indeed bacteremic. The area under the ROC curve gives the probability that the prediction rule assigns a higher likelihood of bacteremia to the patients who are actually bacteremic than to the patients who are not. An ROC area of 1.0 represents perfect discrimination, while an ROC area of 0.5 represents discrimination no better than chance alone. In general, the ROC areas for these prediction rules range from 0.6 to 0.7. A clinically useful model usually has an ROC area of at least 0.7, and some excellent prediction rules have areas as high as 0.85.[16]

The second methodological question is, "How transportable is the model?" One expects the area under the ROC curve to remain stable as the model is tested in different populations. Before a model is accepted into routine clinical practice, it should be validated under several different circumstances. Some investigators use a single data set for both derivation and validation of the model. Here, the data are arbitrarily separated into a training set and a testing set (often in a 2:1 ratio). A more stringent method employs a second data set for validation. Even with this approach, a model validated at the development site might not transport well to other sites.

Several factors potentially impair transportability of prediction models. "Over fitting" causes the model to capture meaningless idiosyncrasies of the training data set.

As one adds more variables to the model, the accuracy continually increases for the training data set, but reaches a peak and then declines for the validation data set. Having at least 10 patients with the least common outcome for each candidate variable limits this problem. Many statistical techniques have been developed to prevent over fitting, but the most important principle centers on using variables selected based on knowledge of the clinical problem, either from previous data or expert opinion. Variables should be included only if they make clinical sense.

In addition, capturing unusual but clinically meaningful features of a population impairs transportation of the model to other populations. For example, the Bates model relied heavily on the predictive power of intravenous drug use and an acute abdominal process. The infrequent occurrence of these diagnoses in the current validation set weakened transportability.

Inclusion of vague endpoints or predictive variables decreases the transportability of the model. For example, the Leibovici model used the presence of chills recorded by several different physicians as abstracted from patient charts. Little effort was made to ensure reliability and accuracy of this subjective variable. Furthermore, this scale used an unvalidated modification of the Karnofsky scale as a subjective determination of functional status.

The inherent heterogeneity of bacteremia also decreases transportability. For example, pneumococcal bacteremia in a patient with pneumonia and gram-negative bacteremia in a patient with pyelonephritis carry different clinical implications. Equating these diagnoses oversimplifies a complex problem.

Finally, different cohort inclusion criteria used to develop the Bates and Leibovici models impair transportability. Leibovici developed his model from a population of patients admitted with a febrile illness. In contrast, Bates developed his model from a population of hospitalized patients who subsequently had blood cultures taken. Again, little effort was made to examine the decision to draw the cultures on these patients. In contrast, this current validation study used a group of patients admitted with suspected infection, not all of whom were febrile. Different predictors of bacteremia in these different patient groups seem logical.

As a third methodological question one must also ask, "How well calibrated is this model?" Calibration reflects the agreement between the results predicted by the model and what is actually observed. In contrast to discrimination, which applies globally to a model, calibration involves sub-groups of patients at varying levels of risk. For example, consider a population where the documented rates of bacteremia for patients of low, intermediate, and high risk are 10%, 30%, and 70%, respectively. A well-calibrated model might predict corresponding bacteremia rates of 8%, 27%, and 72%. In this way, calibration measures how close the performance of an individual model mirrors reality over a range of patient types. To be useful in assisting physicians with medical decision making, a model must demonstrate both good discrimination and calibration.

Next, after consideration of these methodological issues, the most important remaining question is, "How clinically useful is this model?" A clinically useful model must fulfill all the above criteria. In addition, the model should alter the probability of disease or prognosis such that management can be changed. These models fail in this criterion. Based on the posterior probabilities assigned to patients, we would not feel comfortable modifying our clinical decisions about drawing blood cultures or giving empiric antibiotics.

The Yehezkelli paper discusses the problem of clinical utility in detail and it uses sound techniques to assess the validity of the models. While the models did not perform well, the authors provide an excellent blueprint for the validation process. The results of this current study fall nicely in line with previous work. This latest paper emphasizes the need to develop robust models based on specific clinical problems and to demand rigorous validation before clinical application.—**JEROAN J. ALLISON, MD,** *Assistant Professor, Division of General Internal Medicine;* **ROBERT M. CENTOR, MD,** *Professor and Director, Division of General Internal Medicine, University of Alabama at Birmingham, AL*

REFERENCES

1. Yehezkelli Y, Subah S, Elhanan G, Porter RRA, Regev A, Leibovici L. Early prediction of bacteremia: testing of two prediction rules in a university and a community hospital. J Gen Intern Med. 1995; 11:98–103.
2. Bates DW, Cook EF, Goldman L, Lee TH. Predicting bacteremia in hospitalized patients. ACP. 1990;113:495–500.
3. Leibovici L, Greenshtain S, Cohen O, Mor F, Wysenbeek AJ. Bacteremia in febrile patients: a clinical model for diagnosis. Arch Intern Med. 1991;151:1801–6.
4. Capdevila JA, Almirante B, Pahissa A, Planes AM, Ribera E, Martinez-Vazquez JM. Incidence and risk factors of recurrent episodes of bacteremia in adults. Arch Intern Med. 1994;154:411–5.
5. Schwenzer KJ, Gist A, Durbin CG. Can bacteremia be predicted in surgical intensive care unit patients? Intensive Care Med. 1994;20:425–30.
6. Niederman MS, Fein AM. Predicting bacteremia in critically ill patients: a clinically relevant effort? Intensive Care Med. 1994;20: 405–6.
7. Peduzzi P, Shatney C, Sheagren J, et al. Predictors of bacteremia and gram-negative bacteremia in patients with sepsis. Arch Intern Med. 1992;152:529–35.
8. Aube H, Milan C, Blettery B. Risk factors for septic shock in the early management of bacteremia. Am J Med. 1992;93:283–8.
9. Fontanarosa PB, Kaeberlein FJ, Gerson LW, Thomson RB. Difficulty in predicting bacteremia in elderly emergency patients. Ann Emerg Med. 1992;21:7:99–105.
10. Mozes B, Milatiner D, Block C, Blumstein Z, Halkin H. Inconsistency of a model aimed at predicting bacteremia in hospitalized patients. J Clin Epidemiol. 1993;46:1035–40.
11. Pfitzenmeyer P, Decrey H, Auckenthaler R, Michel JP. Predicting bacteremia in older patients. J Am Geriatr Soc. 1995;43:230–5.
12. Mylotte JM, Pisano MA, Ram S, Nakasato S, Rotella D. Validation

of a bacteremia prediction model. Infect Control Hosp Epidemiol. 1995;16:203–9.

13. Ram S, Mylotte J, Pisano M. Rapid classification of positive blood cultures: validation and modification of a prediction model. J Gen Intern Med. 1995;10:82–8.
14. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. N Engl J Med. 1985;313(13):793–9.
15. Centor RM. Signal detectability: the use of ROC curves and their analyses. Med Decis Making. 1991;11:102–6.
16. Goldman L. Cardiac risk in noncardiac surgery: an update. Anesth Analg. 1995;80:810–20.