Validation of Prognostic Models for Melanoma

Phyllis A. Gimotty, PhD,1,2 DuPont Guerry, MD,2,3 and David E. Elder, MB, ChB2,4

Prognostication involves the estimation of the probabilities of clinical outcomes in the history of a disease at some point after the diagnosis is made. A prognostic model characterizes patterns of probabilities and identifies patient-specific factors related to clinical outcomes. Implicit in the definition of a prognostic model is the population for which the probabilities are relevant. Changing the population characteristics can alter the probability of the clinical event. For example, among patients with melanoma, those with thin lesions have a different likelihood of disease recurrence than do those with thicker lesions. Statistical regression models are used to identify factors that explain this heterogeneity.

Multivariate logistic regression and Cox proportional hazards survival analysis are both regression models used to define prognostic factors associated with a poor or good prognosis. These methods are complementary since the former focuses on a clinical event defined in terms of a specific survival time (e.g., melanoma-related death within 5 years of diagnosis) and the latter considers all event rates (e.g., melanoma-related death rates within any period during the follow-up interval), as well as hazard rates (e.g., the likelihood of a clinical event in the next interval given that the patient has survived a number of years). Each approach has its own strengths and weaknesses, and the interpretation of the model depends on adherence to assumptions underlying the method.

Logistic regression can give rise to a model that is convenient to use, as it gives the probability of an outcome at a specific time. However, all patients included in the analysis need to have been observed for the specified time or up to the time of the event. Estimates of probabilities may be unreliable, particularly when many patients are lost to follow-up.

The Cox model is defined more generally. However, it assumes that the hazard associated with the clinical event is the same no matter how long the patient has been followed up. This assumption may be unrealistic in some situations. In contrast with the logistic regression model, which can be used to estimate predicted probabilities, it is not possible to reconstruct group-specific estimates of hazard rates directly from the Cox model coefficients.1 Furthermore, comparisons between different Cox proportional hazards models are problematic because the underlying baseline hazards may differ.

In addition to the choice of analytic strategy, validation of prognostic models is an important step in their development.2 To validate a prognostic model is to demonstrate that the proposed model adequately describes patients other than those whose data were used to develop the model.3 A validation analysis evaluates the extent to which the predicted probabilities from a prognostic model are consistent with the observed outcomes in a new validation data set. There are 2 situations of particular interest.4 If the validation data come from a “similar” population, we are interested in whether the results can be replicated with the new data. If the validation data come from a “different” population, we are interested in whether the results are portable. While the concept is straightforward, the demonstration of the validity of a prognostic model and its implications are more complex.

In this issue of the Journal, the article by Tuthill et al5 presents an important case study for validation of the Clark model,6 a 2-part prognostic model for 8-year survival after diagnosis and surgical therapy of primary melanoma in patients with lesions having only invasive radial growth phase or having vertical growth phase. The model was developed using data from a prospective case series assembled between...
1972 and 1989 by the University of Pennsylvania Pigmented Lesion Clinic (PLC). These unselected patients were without evidence of gross nodal or distant metastases at the time of diagnosis and were seen within 1 year of definitive surgery. This was before the advent of sentinel node biopsies, although some patients had prophylactic node dissections that were positive. Among those with vertical growth phase lesions, 84% of the patients’ outcomes were predicted correctly using a prediction rule that classified those with a predicted probability greater than or equal to .5 to be 8-year survivors and those less than .5 as nonsurvivors. In contrast, the data used in the validation study by Tuthill et al\textsuperscript{5} were derived from patients with localized cutaneous melanoma who participated in a placebo-controlled, randomized trial of vitamin A conducted by the Southwest Oncology Group (SWOG), the results of which were negative. About two thirds of these patients had clinical and pathologic data that were used to develop the prognostic model. All of their tumors were in vertical growth phase; however, unlike those in the Clark cohort, patients had to have lesions at least 0.76 mm in thickness to be eligible for the trial. The ineligibility of patients with lesions less than or equal to 0.75 mm in thickness, the elimination of 27% of those enrolled in the trial because of missing data, and the selectivity associated with the trial-related recruitment process suggest that this cohort is not a “replicate” of the consecutive, clinic-based cohort used to develop the Clark model. Among the patients with vertical growth phase lesions in the SWOG cohort, the outcomes for only 64% were predicted correctly by the Clark model. It is important to understand how these 2 studies differ to address whether and how a model developed for one population can be transported to a different population.

Some statistically significant differences were observed between the 2 cohorts. In the SWOG cohort, there was a higher proportion of males ($P = .019$), a higher proportion of patients whose lesions had absent or nonbrisk tumor infiltrating lymphocytes ($P = .002$), and higher mitotic rates ($P = .003$), suggesting that there was a higher proportion of patients with poorer prognostic factors in the SWOG cohort and that the predicted survival rate might be lower. However, the observed 8-year survival rates were very similar between the 2 cohorts (71% and 69% in the Clark and SWOG cohorts, respectively). Among patients with thicker lesions (>3.6 mm), survival rates were greater for the SWOG cohort than for the Clark cohort (71% and 55%, respectively). These observations suggest that the patients in the SWOG trial were different from those in the Clark cohort and, possibly, that the associations between the prognostic factors and survival differed between the 2 cohorts.

The statistic used to compare the performance of the prognostic models is the “percent correct prediction.” It is important to note that the percentage of patients correctly predicted depends on the case mix of a cohort. To demonstrate this interrelationship, we identified a new validation cohort from the PLC database based on the original eligibility criteria for the Clark model. This new cohort includes 691 patients diagnosed with primary melanoma who also had complete data on all variables in the model (approximately 46% of those seen between 1980 and 1990). In this validation cohort, 40% of these patients had thin lesions, less than 0.75 mm. For the 511 of these lesions with vertical growth phase, we computed predicted probabilities using the original Clark model and used the same decision rule that classified patients with predicted probabilities of .5 or more as survivors. We found that 84% of the patients were classified correctly, exactly as found with the original model. When we then applied the SWOG thickness eligibility criterion, removing all patients in the new PLC validation cohort whose lesion thickness was less than or equal to 0.75 mm (322 vertical growth phase lesions), the proportion of patients with correctly predicted outcomes decreased to 76%. This leads to the speculation that if the mix of patients in the cohort was more like that in the Surveillance, Epidemiology, and End Results registry,\textsuperscript{7} in which approximately 70% of invasive lesions are thin (<1 mm), then the proportion of patients with correctly predicted outcomes might exceed the 84% observed in the Clark study. The percentages of those alive who were predicted to be 8-year survivors and of those dead who were predicted to have died within 8 years were 91% and 66%, respectively, in the new validation cohort compared with 90% and 44%, respectively, in the original Clark model.

Measures such as the percent correct prediction, as well as sensitivity, specificity, and positive and negative predictive value, depend heavily on the distribution of the probabilities for patients in the cohort, which in turn depends on the distribution of characteristics of these patients.\textsuperscript{8} An observed difference in the percent correct prediction may reflect different patient characteristics between the 2 groups rather than instability of the prognostic model. Alternative measures to assess the performance of a prognostic model are worth considering. The Hosmer-Lemeshow statistic is used to compare predicted probabilities from a prognostic model and observed proportions within groups of individuals to assess calibration. Here one looks at groups of individuals with similar predicted probabilities, and when a model is well calibrated, a similar proportion of patients will have been observed to have had the event. On the other hand, the Brier score compares each individual’s predicted probability with 1 and 0 representing those with and those without an event, respectively. A prognostic model with a high degree of accuracy will produce predictive values that are all either close to 1 or to 0. A model’s calibration and accuracy depend on the
strength of the association between the clinical outcome and the prognostic variable.

There have been other articles that addressed validation of the Clark model and have come to different conclusions. For example, Rowley and Cockerell\(^9\) reported that for a cohort of 53 patients, the Clark model correctly predicted outcomes for 59% of the patients, noting that the model was correct 35% of the time in thick (>1.7 mm in thickness) lesions and 69% in thin (<1.7 mm) lesions. In another database of 55 patients, Szymik and Woosley\(^10\) reported that the Clark model correctly predicted outcomes for 85% of the patients with vertical growth phase lesions (n = 47). Clinical and histologic data are reported for all patients in the first study and for only patients with vertical growth phase lesions in the second, making it difficult to compare patients across studies.

It is important to do validation studies including cohorts that are both replicates of the original cohort and cohorts that arise in different contexts, as did Tuthill and colleagues.\(^5\) The ideal prognostic model would encompass all expressions of the disease (eg, in primary melanoma, both growth phases and all thicknesses) and have sufficient information from its prognostic factors to classify patients accurately into those who will have the clinical event and those who will not, in the population in which it is developed. Such a model then should undergo validation in other populations to see whether it can be replicated and whether challenges to its validity are observed when it is used in different contexts. The appropriateness of the use of a prognostic model depends on the intended use. The Clark model does well at identifying those who will survive, especially in patients with invasive radial growth lesions (who have near-perfect disease-free survival) and relatively thin vertical growth phase lesions, and may be useful for protecting patients from too much investigation and therapy. The limitations in predicting those fated to relapse likely can be reduced in the next generation of models with the identification of alternative prognostic factors that are more specific to the clinical event. As with all technologies, the development of a prognostic model is an iterative process ideally subject to continual refinement and evolution. Ultimately, it may be expected that these evolving models will converge toward a common classification scheme. New molecular biomarkers need to be identified through our knowledge of biological processes and evaluated as prognostic factors to improve our ability to accurately and usefully prognosticate.

From the 1Department of Biostatistics and Epidemiology, Division of Biostatistics, 2Melanoma Program, Abramson Cancer Center, 3Department of Medicine, Division of Hematology-Oncology, and 4Department of Pathology and Laboratory Medicine, University of Pennsylvania School of Medicine, Philadelphia.

Supported by SPORE on Skin Cancer grant P50-CA-93372 from the National Cancer Institute, Bethesda, MD.

References