Validation of a Melanoma Prognostic Model

David J. Margolis, MD; Allan C. Halpern, MD; Timothy Rebbeck, PhD; Lynn Schuchter, MD; Raymond L. Barnhill, MD; Judith Fine, BA; Marianne Berwick, PhD, MPH

Background: A “clinically accessible,” 4-variable (patient age, patient sex, tumor location, and tumor thickness) prognostic model has been published previously. This model evaluated variables that were commonly available to the clinician. Because models are heuristic, validity of a prognostic model should be evaluated in a population different from the original population.

Objective: To evaluate the external validity of this 4-variable melanoma prognostic model.

Design: To estimate the external validity of this model, we used a population-based cohort of individuals with melanoma. We also evaluated a 1-variable model (tumor thickness). Estimates of the external validity of these logistic regression models were made using the c statistic and the Brier score.

Settings and Patients: A total of 1261 patients with melanoma evaluated in a multispecialty, university-based practice and 650 patients with melanoma from throughout Connecticut.

Main Outcome Measure: Death from melanoma within 5 years of diagnosis.

Results: The c statistics for the 4-variable model were 0.86 (95% confidence interval [CI], 0.83-0.89) for the university-based practice data set and 0.81 (95% CI, 0.75-0.86) for the Connecticut data set. For thickness alone, the c statistics were 0.83 (95% CI, 0.80-0.86) and 0.79 (95% CI, 0.74-0.85), respectively. Brier scores for the 4-variable model were 0.09 (95% CI, 0.08-0.10) and 0.08 (95% CI, 0.06-0.09) and for the 1-variable model were 0.09 (95% CI, 0.08-0.10) and 0.08 (95% CI, 0.07-0.10), respectively. No significant differences exist between the data sets for the 4- and 1-variable models.

Conclusions: The 4- and 1-variable models are generalizable. The simpler 1-variable model—tumor thickness—can be used with a relatively small loss in accuracy.

Arch Dermatol. 1998;134:1597-1601

Unlike most skin cancers, melanoma contributes significantly to total cancer mortality.1-3 It is a common cause of death by cancer.4 The deaths due to melanoma have been increasing 2% per year for the past 3 decades.2 The incidence rate is estimated to be 11 to 16 per 100 000.3 Patient survival is associated with several prognostic factors.5-11 Of these, probably the most routinely used is tumor thickness, as described by Breslow.10 However, other prognostic factors, such as patient age and sex, site of primary tumor, and histological characteristics of the tumor (eg, Clark level), may also be associated with survival.6,9,12,13 In addition, multivariable prediction models may more accurately predict survival of patients with melanoma than individual prognostic factors.9 However, these models have not gained general acceptance, perhaps because of the difficulties that clinicians and pathologists have in measuring and interpreting the prognostic factors used in these models and the complexity of interpreting a multivariable model.

To create a more user-friendly prediction model for prognosis in melanoma, a model using prognostic factors commonly reported as part of a routine pathology report was proposed by Schuchter et al.6 The covariates in the model were patient age and sex, depth of tumor invasion (thickness), and the anatomical location of the lesion. The original model was established using a cohort of patients with a primary melanoma who were followed up for 10 years. It was reported that this 4-variable model predicted survival better than a model based on tumor thickness alone.

Because all models are heuristic, the validity of a prognostic model should be evaluated in populations different from the original population used in model building. Therefore, we evaluated the external validity of the prognostic model of Schuchter et al6 using a recently reported population-based cohort of individuals with melanoma who had been followed up for 5 years.15 We also compared the predictive ability of this 4-variable model to a 1-variable (tumor thickness) model.
PATIENTS, MATERIALS, AND METHODS

STUDY POPULATION

In 1996, Schuchter et al published a study evaluating 488 patients diagnosed with melanoma as part of the Pigmented Lesion Group (PLG) study. A larger cohort of 1,261 individuals was used in the current study, and the results were compared to those from a smaller cohort of 650 patients identified through the Connecticut Tumor Registry (CT) in 1989. The age, tumor thickness, and tumor location were found to be predictive of patient death from melanoma within 10 years of diagnosis using a multivariable logistic regression model: sex, age, tumor thickness, and tumor location. For the present study, we used a larger cohort (PLG) that includes 1,261 individuals who have been followed up for up to 5 years.

The accuracy factor used by Schuchter et al to assess the 4-variable model was the area under the receiver operator characteristic (ROC) curve. The area was 0.87, which is a good score. The ROC curve is a graphical representation of the true-positive rate (eg, the probability that a positive test result is associated with a death from melanoma—an incorrect response) vs the false-positive rate (eg, the probability that a positive test result is associated with a 5-year survivor—an incorrect response). Unlike a single measurement, such as the percentage of correct responses of a test (eg, a prognostic model), the area under the ROC curve represents all possible responses and is generally believed to be the ideal measure of a test’s ability to discriminate between outcomes.

In 1996, Berwick et al studied a cohort (CT) with melanoma identified through the Rapid Case Ascertainment system of the Cancer Prevention Research Unit for Connecticut at Yale University Medical Center, New Haven. This Rapid Case Ascertainment system is an agent of the Connecticut Tumor Registry, which has functioned as part of the Surveillance, Epidemiology, and End Results program since 1973. The process used to ascertain cases has been described previously. These individuals had been followed up for up to 5 years from the time of melanoma diagnosis. This sample (n = 650) included all those with a validated histopathologic diagnosis of melanoma who lived in Connecticut between January 15, 1987, and May 15, 1989, and whose physician agreed to participate (physician participation rate was 75%). Although this is a population-based cohort, the histopathologic diagnosis and features of melanoma were confirmed and standardized by a single expert pathologist (R.L.B.). A subgroup of this cohort, individuals with localized invasive melanoma, has been used previously to evaluate prognostic factors for death from melanoma within 5 years of diagnosis.

END POINT AND PROGNOSTIC FACTORS

In contrast to the study by Schuchter et al, which evaluated patients for death from melanoma within 10 years of diagnosis, the outcome for this study was death from melanoma within 5 years of diagnosis. In general, determination of death (see the “Comment” section) and the prognostic factors were ascertained in a similar manner in both studies. The prognostic factors chosen for this validation study were based on the study by Schuchter et al: tumor thickness (<0.76, 0.76-1.69, 1.70-3.60, and >3.60 mm), primary lesion location (extremities vs the rest of the body labeled “trunk”), age at diagnosis of melanoma (≤60 or >60 years), and patient sex.

STATISTICAL ANALYSIS

Descriptive statistics were computed for all measured prognostic factors and were reported as the total number of individuals who died of melanoma in the PLG and CT cohorts. The percentage of individuals with a given prognostic factor who died of melanoma within 3 years of diagnosis was also computed. Quantitative differences between those who survived or died within a cohort for a prognostic factor were estimated using Pearson χ² or, if appropriate, χ² for trend.

Logistic regression was used to estimate odds ratios and 95% confidence intervals (CIs). All odds ratio estimates for a prognostic factor were reported as crude odds ratios and as odds ratios adjusted for the other measured prognostic factors.

Multivariable logistic regression was used by Schuchter et al to formulate the clinically friendly 4-variable prediction model for 10-year survival. This technique was also used in the present study to create a prediction for 5-year survival. In addition, a model was created using only Breslow depth of tumor invasion (ie, tumor thickness, <0.76, 0.76-1.69, 1.70–3.60, and >3.60 mm). The performance of each model was evaluated by the goodness-of-fit test (calibration) and discrimination. Brier scores were estimated as a measure of goodness-of-fit. Model discrimination was estimated by the c statistic, which is equivalent here to the area under the ROC curve. This is a widely used estimate of discrimination and is presented as the probability that 1 individual as compared to another will survive 5 years. Quantitative differences between Brier scores or c statistics for the 4- and 1-variable models were evaluated using a z statistic test. With the exception of melanoma prognostic modeling, these test statistics have rarely been used to evaluate dermatologic ailments.

In summary, a 4-variable prognostic model and a 1-variable prognostic model were created using the PLG and CT data sets. All statistical computations were performed using a software program (Stata version 5, Stata Corp, College Station, Tex), except for calculation of the z statistic, which was done manually.

RESULTS

One hundred sixty-seven (13.2%) of 1,261 individuals in the PLG cohort died of melanoma within 5 years of diagnosis. Death from melanoma was significantly associated with individuals who had lesions on their trunk, who were male, who were older than 60 years, and who had thicker lesions (Table 1 and Table 2). There were no differences in the inferences from the analysis of either crude or adjusted odds ratios (Table 2).

Eighty (12.3%) of 650 individuals in the CT cohort died of melanoma within 5 years of diagnosis. Death from
melanoma was significantly associated with individuals who had thicker lesions (Tables 1 and 2). In contrast to the PLG data set, death from melanoma was not associated with lesion location, patient age, or patient sex (Tables 1 and 2). Except for patient sex, crude and adjusted odds ratio estimates for the association between prognostic factors and death from melanoma for tumor thickness were similar.

Using the 4-variable model, the c statistic was 0.86 (95% CI, 0.83-0.89) for the PLG data set and 0.81 (95% CI, 0.75-0.86) for the CT data set (Table 3). Using the 1-variable model (tumor thickness alone), the c statistic was 0.83 (95% CI, 0.80-0.86) for the PLG data set and 0.79 (95% CI, 0.74-0.85) for the CT data set. Brier scores for the 4-variable model were 0.09 (95% CI, 0.08-0.10) for the PLG data set and 0.08 (95% CI, 0.06-0.09) for the CT data set. Brier scores for the 1-variable model (thickness alone) were 0.09 (95% CI, 0.08-0.10) for the PLG data set and 0.08 (95% CI, 0.07-0.10) for the CT data set (Table 3).

Therefore, at 5-year follow-up, the accuracy of the 4- or 1-variable model in predicting prognosis, as estimated by the Brier score and c statistic, was not significantly different. This was true when 1- and 4-variable models were compared using the PLG or CT data sets. With 1 exception, no statistically significant differences (P > .10) between Brier scores or c statistics were noted for any combination of models or populations. In 1 case, the P values comparing the accuracy of the 4- and 1-variable models within the PLG data set were .06 for the c statistic but greater than .10 for the Brier score.

COMMENT

Prognostic models are intended to predict the probability of an outcome (eg, disease or death) in a specific data set. The goal of this process is to develop an accurate model. Accuracy is the degree to which the predicted probability of an event agrees with the actual observed outcome in the data set. Ultimately, however, this heuristic approach is too simplistic because models are seldom created for use only in the original modeling data set (ie, for this study, the PLG data set).

A problem with the use of prognostic models is that the external validity or generalizability of a model cannot be described in conclusive terms. However, if a prognostic model is to be appropriately clinically applied in a different population of inference, its performance must be tested in an external (validation) data set. Externally refers to the extent to which the results of the model represent events seen in a referent population of interest. Lack of external validity cannot be corrected statistically; it is essential for a clinician to evaluate and understand the generalizability of a prognostic model to his or her own clinical setting.

Estimates of generalizability can be determined by evaluating the accuracy of a predictive model in different clinical settings. Although no consensus exists as to how a model should be validated, these assessments are often classified in terms of calibration (goodness of fit) and discrimination.

Calibration is the degree to which the predicted probability agrees with the actual event. Discrimination is the degree to which a prediction from the model can separate those who will have the outcome from those who will not. For example, a well-calibrated model can correctly predict that an individual has a 60% chance of dying of melanoma within 5 years of diagnosis, whereas a discriminating model correctly distinguishes between who will die of melanoma within 5 years of diagnosis and who will not. When models lack validity, it is because the predictions do not differentiate among those who will have the outcome and those who will not (poor discrimination) or because predictions from a model do not estimate the average rate of the outcome experienced by an individual within a particular subgroup (lack of calibration). Therefore, the generalizability of a model is best estimated by determining both calibration and discrimination using data from many sites.

Many statistical techniques exist to measure a model’s goodness of fit. A commonly used technique in epidemiological studies is the Hosmer-Lemeshow statistic, a modification of a Pearson χ² test. It has not, however, been used to make comparisons between studies, and the estimates of calibration of this test are sensitive to sample size. For this reason, in the present study, we chose to use the Brier score. Brier scores are commonly used by atmospheric scientists to summarize the forecast performance of their models and have been used by epidemiologists.

A Brier score is the average of the mean squared error of the predicted and the observed event for any data set. Scores can vary between 0 and 1. A score of 0 indicates that the predictions from a model are perfect. A score of 1 indicates that the predictions are no better than chance. A score of 2 indicates that the predictions are twice as likely to be correct as chance. A score of 2.5 indicates that the predictions are twice as likely to be correct as chance. These scores are not directly comparable across different data sets. A score of 0.25 indicates that the predictions are twice as likely to be correct as chance. A score of 0.75 indicates that the predictions are twice as likely to be correct as chance. A score of 1.0 indicates that the predictions are twice as likely to be correct as chance. A score of 1.5 indicates that the predictions are twice as likely to be correct as chance. A score of 2.0 indicates that the predictions are twice as likely to be correct as chance. A score of 2.5 indicates that the predictions are twice as likely to be correct as chance. A score of 3.0 indicates that the predictions are twice as likely to be correct as chance. A score of 3.5 indicates that the predictions are twice as likely to be correct as chance. A score of 4.0 indicates that the predictions are twice as likely to be correct as chance. A score of 4.5 indicates that the predictions are twice as likely to be correct as chance. A score of 5.0 indicates that the predictions are twice as likely to be correct as chance.

To evaluate discrimination, we used a c statistic, which for the type of models presented in this study is equivalent to the area under the ROC curve. This is a common approach that is widely used to evaluate discrimination. The c index is calculated by constructing a

Table 1. Individuals With a Prognostic Factor Who Died of Melanoma

<table>
<thead>
<tr>
<th>Prognostic Factor</th>
<th>Patients, No. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PLG</td>
</tr>
<tr>
<td></td>
<td>CT</td>
</tr>
<tr>
<td>Trunk (head, neck, and trunk)</td>
<td>136 (17.5)†</td>
</tr>
<tr>
<td></td>
<td>59 (13.8)</td>
</tr>
<tr>
<td>Extremities</td>
<td>31 (6.4)</td>
</tr>
<tr>
<td></td>
<td>21 (9.4)</td>
</tr>
<tr>
<td>Men</td>
<td>106 (17.4)†</td>
</tr>
<tr>
<td></td>
<td>34 (11.1)</td>
</tr>
<tr>
<td>Women</td>
<td>61 (9.4)</td>
</tr>
<tr>
<td></td>
<td>46 (13.4)</td>
</tr>
<tr>
<td>Age, y</td>
<td></td>
</tr>
<tr>
<td>≤60</td>
<td>95 (9.8)†</td>
</tr>
<tr>
<td></td>
<td>43 (11.0)</td>
</tr>
<tr>
<td>&gt;60</td>
<td>72 (25.0)</td>
</tr>
<tr>
<td></td>
<td>37 (14.2)</td>
</tr>
<tr>
<td>Tumor thickness, mm</td>
<td></td>
</tr>
<tr>
<td>&lt;0.76</td>
<td>8 (1.5)‡</td>
</tr>
<tr>
<td></td>
<td>5 (1.8)‡</td>
</tr>
<tr>
<td>0.76-1.69</td>
<td>28 (7.6)</td>
</tr>
<tr>
<td></td>
<td>14 (7.8)</td>
</tr>
<tr>
<td>1.70-3.60</td>
<td>74 (25.6)</td>
</tr>
<tr>
<td></td>
<td>25 (24.8)</td>
</tr>
<tr>
<td>&gt;3.60</td>
<td>57 (20.0)</td>
</tr>
<tr>
<td></td>
<td>20 (32.8)</td>
</tr>
</tbody>
</table>

*P < .01; †P < .05; ‡P < .001; †P < .01 (Pearson χ² for trend).

©1998 American Medical Association. All rights reserved.
than 0.9 can be thought of as excellent. A higher than 0.8 can be thought of as good, and higher than 0.5 being no predictive value, and 0 being a perfect negative predictive value. A perfect positive predictive value is tied are scored as 0.5, and pairs for which the prognostic score is greater for the patient with the positive outcome are scored as 1. Pairs for which the prognostic score is less than 0.76 are scored as 0.16. The c index is the total score given a score of 1, pairs for which the prognostic score is tied are scored as 0.5, and pairs for which the prognostic score is less than 0.76 are scored as 0.16. The mathematical expression for the 1-variable model is logit (probability of death from melanoma within 5 years of diagnosis) = 5.25 − 0.66 (age 60 y) − 0.23 (male) − 3.31 (lesion depth 0.76-1.69 mm) = 5.13 (lesion depth 1.70-3.60 mm) − 1.11 (lesion depth >3.60 mm). The mathematical expression for the 4-variable model is logit (probability of death from melanoma within 5 years of diagnosis) = 4.18 + 1.68 (lesion depth 0.76-1.69 mm) − 3.31 (lesion depth 1.70-3.60 mm) − 4.18 (lesion depth >3.60 mm). Estimates are reported with 95% confidence intervals estimated by the bootstrap technique on 1000 samples. PLG indicates pigmented lesion group; CT, Connecticut cohort with melanoma.

Important differences do, however, exist between the data sets (Tables 1 and 2). One potential explanation of these differences may be selection bias. For example, 25% of individuals older than 60 years died of melanoma in the PLG data set but only 14% of individuals older than 60 years died of melanoma in the CT data set. In the CT data set, only 75% of eligible patients were interviewed and, therefore, initially entered into the CT data set. A comparison of those who were not interviewed (because of death, physician refusal, or patient refusal often because of illness) shows that they were slightly older and had slightly thicker tumors than the interviewed patients. The magnitude and direction of this bias on the prognostic factors is unknown. As a result, the restriction of cases to “death from melanoma” might have been too strict, and additional evaluations using all-cause mortality should be conducted. Finally, there may be important differences among the individuals who choose to be cared for in a specialty center like the PLG and those who live in Connecticut. For example, individuals with a family history of melanoma might seek care out of state in a PLG-like practice.

A limitation of this study is the 5 years of follow-up. Although most people may die of melanoma within 5 years of diagnosis, many still die of melanoma later than 5 years after diagnosis. The original model of Schuchter et al using the PLG data set followed up patients for 10 years. When available, data sets with 10 years of follow-up after diagnosis should be used to fully evaluate the 4- and 1-variable models. Of the variables studied, tumor thickness may be the only variable required to predict the chance of death from melanoma within 5 years of diagnosis, and that the additional variables are needed to accurately predict death from melanoma within 10 years of diagnosis.

<table>
<thead>
<tr>
<th>Prognostic Factor</th>
<th>Crude</th>
<th>Adjusted</th>
<th>Crude</th>
<th>Adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truncal lesion</td>
<td>3.10 (2.06-4.66)</td>
<td>3.02 (1.90-4.79)</td>
<td>1.55 (0.92-2.63)</td>
<td>1.92 (0.99-3.73)</td>
</tr>
<tr>
<td>Age, &gt;60 y</td>
<td>3.08 (2.19-4.33)</td>
<td>1.93 (1.30-2.68)</td>
<td>1.33 (0.83-2.13)</td>
<td>1.09 (0.62-1.93)</td>
</tr>
<tr>
<td>Men</td>
<td>2.04 (1.46-2.86)</td>
<td>1.26 (0.84-1.87)</td>
<td>0.80 (0.50-1.29)</td>
<td>1.24 (0.69-2.25)</td>
</tr>
<tr>
<td>Tumor thickness, mm</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;0.76</td>
<td>1.00 (Reference)</td>
<td>1.00 (Reference)</td>
<td>1.00 (Reference)</td>
<td>1.00 (Reference)</td>
</tr>
<tr>
<td>0.76-1.69</td>
<td>5.36 (2.42-11.91)</td>
<td>5.57 (2.50-12.44)</td>
<td>4.43 (1.57-12.52)</td>
<td>4.32 (1.52-12.24)</td>
</tr>
<tr>
<td>1.70-3.60</td>
<td>27.38 (12.94-57.92)</td>
<td>25.20 (11.83-53.70)</td>
<td>17.70 (6.55-47.79)</td>
<td>18.04 (6.65-48.95)</td>
</tr>
<tr>
<td>&gt;3.60</td>
<td>65.12 (29.59-143.34)</td>
<td>56.76 (25.40-126.83)</td>
<td>26.24 (9.34-73.78)</td>
<td>26.63 (9.28-76.44)</td>
</tr>
</tbody>
</table>

*Per Schuchter et al, adjusted odds ratios were adjusted for all 4 variables. CI indicates confidence interval; PLG, pigmented lesion group; and CT, Connecticut cohort with melanoma.

Table 2. Odds Ratios for Death From Melanoma Within 5 Years of Diagnosis for the Crude and Adjusted Prognostic Factors

Table 3. Brier Score and c Statistic Estimates for the 4-Variable and 1-Variable Models Predicting Death From Melanoma Within 5 Years of Diagnosis

*The mathematical expression for the 4-variable model is logit (probability of death from melanoma within 5 years of diagnosis) = 5.25 − 0.66 (age >60 y) − 0.23 (male) − 3.31 (lesion depth 0.76 - 1.69 mm) − 3.23 (lesion depth 1.70 - 3.60 mm) − 1.11 (lesion depth >3.60 mm). The mathematical expression for the 1-variable model is logit (probability of death from melanoma within 5 years of diagnosis) = 4.18 + 1.68 (lesion depth 0.76-1.69 mm) − 3.31 (lesion depth 1.70-3.60 mm) − 4.18 (lesion depth >3.60 mm). Estimates are reported with 95% confidence intervals estimated by the bootstrap technique on 1000 samples. PLG indicates pigmented lesion group; CT, Connecticut cohort with melanoma.
In summary, our results show that the models created from the PLG data set are generalizable to the CT population. The simpler 1 variable—tumor thickness—can be used with a relatively small loss in accuracy. However, deaths caused by melanoma do occur more than 5 years after diagnosis. Therefore, when data sets with 10 years of follow-up become available, both the 1- and 4-variable models should be reevaluated. In the future, more accurate and generalizable multivariable models will likely include among their prognostic factors molecular and biological attributes of the primary tumor (eg, mutations in oncogenes and measures of angiogenesis) and markers of metastatic capacity (eg, nodal staging and assessment of blood for messenger RNA for tyrosinase).

Accepted for publication July 23, 1998.
This work was funded partially by grants AG 00715, CA 42101, and CA 75434 from the National Institutes of Health, Bethesda, Md, and a Biomedical Research Support grant from the Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, Conn.

Presented in part as an oral abstract at the International Dermatopathology Association meeting, Orlando, Fla, February 26, 1998.

We thank the following institutions, without whose assistance this research would not have been possible: Connecti-cut Dermatopathology Laboratory Inc, Laboratory of Hope-Ross & Portenoy, University of Connecticut Dermatopathology Laboratory, Yale Dermatopathology Laboratory, Hartford Hospital, Yale-New Haven Hospital, St Francis Hospital & Medical Center, Bridgeport Hospital, Waterbury Hospital, Hospital of St Raphael, Danbury Hospital, New Britain General Hospital, Norwalk Hospital, St Vincent's Medical Center, The Stamford Hospital, Middlesex Hospital, Mt Sinai Hospital, St Mary's Hospital, Lawrence & Memorial Hospital, Manchester Memorial Hospital, Greenwich Hospital Association, Veterans Memorial Medical Center, Griffin Hospital, Bristol Hospital, St Joseph Medical Center, UCONN Health Center/John Dempsey Hospital, William W Backus Hospital, Park City Hospital, Charlotte Hungerford Hospital, Windham Community Memorial Hospital, Milford Hospital, Day Kimball Hospital, Rockville General Hospital, Bradley Memorial Hospital, The Sharon Hospital, New Milford Hospital, Johnson Memorial Hospital, Winstead Memorial Hospital, Welsey (Rhode Island) Hospital and the Pigmented Lesion Group, and the University of Pennsylvania Medical Center. In addition, we thank Dupont Guerry, MD, for his editorial assistance and Breslow A. Thickness, cross-sectional area and depth of invasion in the prognosis of cutaneous melanoma. Ann Surg. 1970;172:902-908.


Murphy AH. What is a good forecast? an essay on the nature of goodness in weather forecasting. Weather Forecast. 1993;8:281-293.


