The Performance of SolarScan

An Automated Dermoscopy Image Analysis Instrument for the Diagnosis of Primary Melanoma

Scott W. Menzies, MB, BS, PhD; Leanne Bischof, M Biomed E; Hugues Talbot, PhD; Alex Gutenev, PhD; Michelle Avramidis, BSc; Livian Wong, BSc; Sing Kai Lo, PhD; Geoffrey Mackellar, PhD, BSc; Victor Skladnev, PhD; William McCarthy, MB, BS, MEd; John Kelly, MD, BS; Brad Cranney, MB, BS; Peter Lye, MB, BS; Harold Rabinovitz, MD; Margaret Oliviero, ARNP; Andreas Blum, MD; Alexandra Virol, B Med; Brian De’Ambrosis, MB, BS; Roderick McCleod, MB, BS; Hiroshi Koga, MD; Caron Grin, MD; Ralph Braun, MD; Robert Johr, MD

Objective: To describe the diagnostic performance of SolarScan (Polartechnics Ltd, Sydney, Australia), an automated instrument for the diagnosis of primary melanoma.

Design: Images from a data set of 2430 lesions (382 were melanomas; median Breslow thickness, 0.36 mm) were divided into a training set and an independent test set at a ratio of approximately 2:1. A diagnostic algorithm (absolute diagnosis of melanoma vs benign lesion and estimated probability of melanoma) was developed and its performance described on the test set. High-quality clinical and dermoscopy images with a detailed patient history for 78 lesions (13 of which were melanomas) from the test set were given to various clinicians to compare their diagnostic accuracy with that of SolarScan.

Setting: Seven specialist referral centers and 2 general practice skin cancer clinics from 3 continents. Comparison between clinician diagnosis and SolarScan diagnosis was by 3 dermoscopy experts, 4 dermatologists, 3 trainee dermatologists, and 3 general practitioners.

Patients: Images of the melanocytic lesions were obtained from patients who required either excision or digital monitoring to exclude malignancy.

Main Outcome Measures: Sensitivity, specificity, the area under the receiver operator characteristic curve, median probability for the diagnosis of melanoma, a direct comparison of SolarScan with diagnoses performed by humans, and interinstrument and intrument reproducibility.

Results: The melanocytic-only diagnostic model was highly reproducible in the test set and gave a sensitivity of 91% (95% confidence interval [CI], 86%-96%) and specificity of 68% (95% CI, 64%-72%) for melanoma. SolarScan had comparable or superior sensitivity and specificity (85% vs 65%) compared with those of experts (90% vs 59%), dermatologists (81% vs 60%), trainees (85% vs 36%; P=0.06), and general practitioners (62% vs 63%). The intraclass correlation coefficient of instrument repeatability was 0.86 (95% CI, 0.83-0.88), indicating an excellent repeatability. There was no significant interinstrument variation (P=0.80).

Conclusions: SolarScan is a robust diagnostic instrument for pigmented or partially pigmented melanocytic lesions of the skin. Preliminary data suggest that its performance is comparable or superior to that of a range of clinician groups. However, these findings should be confirmed in a formal clinical trial.

Arch Dermatol. 2005;141:1388-1396

Although early detection of melanoma is critical for controlling mortality from the disease, it is clear that diagnostic accuracy in the field is suboptimal.1,2 Therefore, a considerable effort has gone into producing automated diagnostic instruments (so-called machine diagnosis) for primary melanoma of the skin. Studies conducted before March 20023 and after March 20024-11 were reviewed; from these reviews, basic quality requirements for describing such instruments were outlined: (1) selection of lesions should be random or consecutive; (2) inclusion and exclusion criteria should be clearly stated; (3) all lesions clinically diagnosed as melanocytic should be analyzed; (4) the study setting should be clearly defined; (5) to avoid verification bias, clearly benign lesions that were not excised should be included, with the diagnostic gold standard being short-term follow-up with digital monitoring; (6) instrument calibration should be reported; (7) repeat-
ability analysis should be carried out (interinstrument and intraintstrument); (8) classification should be carried out on an independent test set; and (9) computer diagnosis should be compared with human diagnosis.

We have previously published12 pilot data on an automated diagnostic instrument (Mk1 Skin PolarProbe; Polartechtics Ltd, Sydney, Australia), which uses image analysis of dermoscopy (surface microscopy) features of pigmented skin lesions. Following that report, the digital surface microscopy (dermoscopy) video instrument SolarScan (Polartechtics Ltd) was developed, and data were collected from 9 clinical sites around the world. Herein we assess the performance of this instrument in terms of these quality requirements.

### METHODS

#### DATA COLLECTION

Between the period of June 15, 1998, and September 30, 2003, images were taken of pigmented skin lesions using SolarScan from 9 clinical centers. Of these, 7 were specialist referral centers: the Sydney Melanoma Unit (Sydney Melanoma Diagnostic Centre), Sydney, Australia; Skin and Cancer Associates, Miami, Fla; Department of Dermatology, University of Tubingen, Tubingen, Germany; the Skin and Cancer Foundation, Sydney; KellyDerm, the private clinic of one of the authors (J.K.), Melbourne, Australia; and South East Dermatology and the Princess Alexandra Hospital, Brisbane, Australia. Two centers were at private skin cancer clinics in Australia: Central Coast Skin Cancer Clinic, Gosford, and the Chatswood Skin Cancer Clinic, Sydney, all staffed by general practitioners. Images were taken after formal written consent by patients, and the research protocol was reviewed by the local ethics committee of each clinic site.

The instrument specifications of the SolarScan have been described previously.13 In addition to imaging, a patient history was recorded that indicated whether the lesion had, within the previous 2 years, bled without being scratched, changed in color or pattern, or increased in size (answer choices: yes, no, uncertain). In all but 1 clinic site, the sole indication for imaging was that the pigmented lesion was to be excised, usually because of a clinical suspicion. However, clinics were inconsistent in imaging excised lesions from their own practices, with some clinics obtaining images of lesions with a predominately high probability of melanoma. Reports of histopathologic findings provided by clinicians were then used as the gold standard for diagnosis. These lesions made up 71% of the data set. In 1 clinic site (Sydney Melanoma Unit), some images were taken of nonmelanocytic pigmented lesions that were diagnosed clinically but not excised. These lesions represented only 3% of the total image set. Also at the Sydney Melanoma Unit, melanocytic lesions that underwent short-term digital monitoring over a 3-month period and remained unchanged were classified as benign according to the previously described protocol.13 These lesions were either moderately atypical melanocytic lesions without a patient history of change or mildly atypical lesions with a history of change. These images represented 26% of the data set. In all centers, some repeated images were taken to permit a reproducibility analysis.

Lesions were excluded from analysis if they were outside the field of view (24 X 18 mm), could not be calibrated reliably because of contamination of calibration surfaces, or had excess artifacts (hair, air bubbles, or movement artifacts). Clipping excess hair before imaging was suggested. Lesions that were nonpigmented, ulcerated, or at an acral site, or that were diagnosed as pigmented basal cell carcinoma, pigmented Bowen disease, or squamous cell carcinoma were also excluded. Although pure amelanotic lesions were excluded (using dermoscopy imaging of absent brown, blue, gray, or black pigmentation), partially pigmented or lightly pigmented lesions were included. Finally, lesions from anatomical areas that could not be imaged adequately using the SolarScan headpiece (eg, eyelids, some parts of the pinna, some genital sites, and perianal and mucosal surfaces) were unable to be assessed. The diagnostic frequency of the 2430 analyzed lesions is shown in Table 1.

#### IMAGE PROCESSING

Each image was calibrated using a procedure of black and white shading, shading correction, setup of camera dynamic range, and capture of an image of a reference surface of known reflectivity, followed by tracing of the colors of the captured lesion to a color space common for all SolarScan instruments, as previously described13 (System and Method for Examining, Recording and Analyzing Dermatological Conditions; US Patent filing No. 09/473270). The lesion border was then determined by a semiautomated procedure and confirmed as accurate by 2 clinicians (S.W.M. and H.K.). For those lesions in which the border was not correctly segmented by this procedure (24%), the lesion border was manually created. An automated procedure was then performed to mask out hair and air bubble artifacts. A total of 103 automated image analysis variables consisting of various properties of color, pattern, and geometry were extracted from the segmented lesion images (Diagnostic Feature Extraction in Dermatological Examination; US Patent filing No. 10/478078).

#### ALGORITHM DEVELOPMENT

The entire set of 2430 lesions was divided into a training set and an independent test set at a ratio of approximately 2:1, respectively. These sets were created by a random allocation of lesions stratified by diagnostic category and Breslow thickness. Before algorithm development, each lesion diagnostic category was assigned a “weight” based on a linear representation (range,
Two sets of repeated images were used to test the intrument reproducibility of the diagnostic algorithm. First, repeated images with an orientation of 90° rotation were taken of 387 lesions. Second, 304 images of lesions that were undergoing 3-month digital monitoring and that remained unchanged were collected and compared with their baseline image taken 3 months before. These were taken at the same orientation. In both of these sets, the images were processed as described herein and the algorithm probability calculated. The intraclass correlation coefficient (ICC) (3,1) was used to assess the intramachine reliability. Here, a coefficient greater than 0.75 indicates excellent reliability. We also denoted the reproducibility by describing the median of the algorithm probability differences between the repeated images and median experimental error. Here, the experimental error equals the difference between repeated lesion probabilities times 100 divided by the lesion probability. Finally, the repeatability of the algorithm diagnosis using the arbitrary index cutoff (ie, the percentage of lesions that have the same diagnosis in their repeats) for both true melanomas or nonmelanomas were described.

**Interinstrument Reproducibility**

A total of 48 lesion images were taken on 3 SolarScan instruments (3 repeated images per instrument). The images were processed, the algorithm probabilities calculated, and the mean value of the repeats given. The ICC (2,1) was used to assess the intermachine reliability. Again, a coefficient greater than 0.75 indicates excellent reliability. In this experimental design, the calculated interinstrument experimental percentage error is the addition of the intramachine error and the true interinstrument percentage errors. Hence, the true interinstrument error can be calculated. For this study, the experimental percentage error was the standard error of the mean (repeats) times 100 divided by the mean lesion probability.

**DIAGNOSIS BY HUMANS VS ALGORITHM**

To assess performance of the SolarScan diagnostic melanocytic algorithm vs diagnoses performed by humans, all melanocytic lesions from the independent test set taken at the Sydney Melanoma Unit that had clinical and dermoscopy photographic images (taken with a Heine Dermaphot camera, Heine Ltd, Herrings, Germany); patient details of age, sex, and lesion site; and a recorded history of whether the lesion had, within the past 2 years, bled without being scratched, changed in color or pattern, or increased in size (answer choices: yes, no, uncertain) were collected. All lesions had diagnoses based on histological findings. This resulted in a set of 78 melanocytic lesions (Table 2). These images and patient histories were given to 13 independent clinicians who were not involved in the data collection for the study. Three were international dermoscopy experts who headed pigmented le-
sion clinics (C.G., R.B., and R.J.), 4 were practicing dermatologists from the Sydney metropolitan area, 3 were dermatology registrars (trainee dermatologists), and 3 were primary care physicians from the Sydney metropolitan area. For each of these lesions, the following questions were answered: diagnosis of (1) melanoma (in situ or invasive) or (2) benign nevus (including dysplastic); probability of melanoma (0%-100%), where 0% is certain for being benign and 100% represents certain melanoma; management by (1) excision or referred for a second opinion, (2) close observation (e.g., monitoring for 3 months), or (3) routine observation.

RESULTS

ALGORITHM PERFORMANCE
DISTINGUISHING MELANOMAS
FROM ALL BENIGN PIGMENTED LESIONS

From the training set of 1644 lesions, of which 260 were melanomas (97 in situ and 163 invasive; overall median Breslow thickness, 0.37 mm), a diagnostic algorithm was developed to distinguish melanomas from all benign pigmented lesions. This model was run on an independent test set of 786 lesions, 122 of which were melanomas (47 in situ and 75 invasive; overall median Breslow thickness, 0.36 mm) (see the "Methods" section and Table 1). The receiver operator characteristic curve of both diagnostic models is shown in Figure 1A. Here, the performance of the algorithm is shown to be reproducible, with little difference of the area under the receiver operator characteristic between the test and training set (0.871 vs 0.877, respectively; \( P = .78 \) for 2-sided Z test). Using an arbitrary cutoff developed in the training set, the sensitivity for melanoma was 90% (95% confidence interval [CI], 86%-94%) and specificity 61% (95% CI, 58%-64%). In the test set, this was shown to be reproducible with a sensitivity of 91% (95% CI, 86%-96%) and a specificity of 65% (95% CI, 61%-69%). On examination of the algorithm performance as a function of diagnostic categories, no difference existed in the proportion of correctly classified lesions in the training or test set (Table 3). However, although the algorithm performed well on melanocytic lesions, it performed poorly on benign nonmelanocytic lesions. In particular, seborrheic keratoses that were diagnosed on routine dermoscopy examination were correctly classified by the algorithm in only 6 (13%) of 47 cases (combined test and training sets). In addition, hemangiomas and dermatofibromas were correctly classified in less than 50% of cases.

ALGORITHM PERFORMANCE
DISTINGUISHING MELANOMA
FROM BENIGN MELANOCYTIC LESIONS

Because the developed algorithm failed to adequately distinguish melanomas from pigmented nonmelanocytic lesions, a new algorithm was developed to distinguish melanomas from benign melanocytic lesions. Here, the training set consisted of 260 melanomas and 1239 benign nonmelanocytic lesions, and the test set, 122 melanomas and 596 benign nonmelanocytic lesions, as detailed in Table 3. The median Breslow thickness was 0.37 mm. The optimum model remained that described all pigmented lesions. Figure 1B shows the receiver operator characteristic curves of both diagnostic models. The area under the curve is larger than the algorithm for modeling all pigmented lesions (Figure 1A), and again, there is good reproducibility be-
between the performance of the algorithm in the test and training set (0.881 vs 0.887 receiver operator characteristic curve areas, respectively; \( P = .77 \) for 2-sided Z test). Using an arbitrary cutoff developed in the training set, the sensitivity for melanoma was 90% (95% CI, 86%-94%) and specificity was 64% (95% CI, 61%-67%). In the test set, this result was shown to be reproducible with a sensitivity of 91% (95% CI, 86%-96%) and a specificity of 68% (95% CI, 64%-72%). The model performance as a function of diagnosis category is described in Table 3. As stated, there was excellent reproducibility in the test set, with no significant difference existed between the training and test sets in the percentages of correctly classified lesions as a function of the histological diagnosis using the \( \chi^2 \) test (with or without Bonferroni adjustment of \( \alpha \) level).

Rather than expressing the algorithm classifier as diagnostic in the test and training sets, the number and percentage of algorithm true-positive diagnoses are shown in bold type. No significant difference existed between the training and test sets in the percentages of correctly classified lesions as a function of the histological diagnosis using the \( \chi^2 \) test (with or without Bonferroni adjustment of \( \alpha \) level).

The algorithm was weighted to preferentially detect thicker melanomas over thinner lesions (see the “Methods” section). In this regard, a significant difference existed in the Breslow thicknesses between the true positive (correctly classified) melanomas (median, 0.4 mm) vs those misclassified (median in situ) in the training set. However, this difference failed to reach significance in the independent test set (Table 4). Similarly, a significant difference existed in the mean algorithm probability of melanoma between in situ lesions, invasive melanomas thinner than 1 mm, and lesions at least 1-mm thick in the training set (\( P < .001 \)). Again, this difference failed to reach significance in the test set (\( P = .13 \); Kruskal-Wallis test) (Table 5).

**INTRAMACHINE AND INTRAINSTRUMENT REPRODUCIBILITY**

The intramachine reproducibility was analyzed in 2 ways. First, repeated images were taken of 387 melanocytic lesions, with different orientations on the same instrument, and the algorithm (probability) assessed (see the
“Methods” section). The ICC (3,1) was 0.86 (95% CI, 0.83-0.88), which indicates an excellent correlation. The median absolute difference of the probabilities between the repeated images was 1.2%, with a median melanoma probability of the lesion set of 12%. The median experimental error was 7.6%. These errors have been plotted as a function of algorithm index in Figure 3. Finally, the algorithm diagnosis reproducibility was 95% for true melanomas and 83% for true benign melanocytic lesions.

Second, repeated images of 304 lesions were taken 3 months after the baseline images using the same instrument. All of these were morphologically unchanged and hence benign. The ICC (3,1) was 0.73 (95% CI, 0.67-0.78). The median absolute difference of the probabilities between the repeated images was 0.14%, with a median melanoma probability of the lesion set of 2.9%. The median experimental error was 4.4%. Finally, the algorithm diagnosis reproducibility was 84% (all true benign melanocytic lesions).

To assess whether there was any effect of having a lesion border generated by the manual or automated method (see the “Methods” section), we analyzed 22 paired lesion images taken on the same instrument with an automated border generated on one and a manual border on the other. The ICC (3,1) was 0.89 (95% CI, 0.74-0.95), which indicated an excellent correlation.

We examined the interinstrument reproducibility of the algorithm (algorithm probability) by analyzing 48 pigmented lesions on 3 SolarScan instruments. The ICC (2,1) was 0.88 (95% CI, 0.82-0.93), well above the 0.75 limit of excellent reliability. There was no significant difference between the experimental percentage errors of the interinstrument (11.4%) and intrument (11.8%) reproducibility (P = .13, Wilcoxon signed rank test). This indicates that no significant true interinstrument variation exists among the 3 instruments.

THE DIAGNOSTIC PERFORMANCE OF HUMANS VS SOLARSCAN

To test the performance of the diagnostic algorithm (for melanocytic lesions only), all lesions that had good-quality clinical and dermoscopy images and complete patient and lesion history details from the independent test set were collected from 1 site (Sydney Melanoma Unit) and compared with a range of clinician groups (see the “Methods” section and Table 2). When we compared the diagnoses performed by humans with those of the SolarScan algorithm (based on the index cutoff as described herein), no statistically significant difference existed in the sensitivity (based on either the absolute diagnosis or the decision to excise as diagnosing melanoma) between any clinician group and the algorithm (Table 6). However, a significant power problem secondary to a low sample size of melanomas may confound these results. In this regard, SolarScan had a sensitivity comparable with that of dermoscopy experts, dermatologists, and trainee dermatologists, and had a substantial superior sensitivity compared with general practitioners. For analysis of specificity, SolarScan’s performance was superior to that of trainee dermatologists (P = .01) and had a higher specificity than all 4 clinical groups (based either on absolute diagnosis or the decision to not excise a benign lesion).

On the assumption that the prevalence of melanoma was the same in the clinical test as in the population of excised lesions in the field, the positive predictive value
Numerous systems that automatically diagnose pigmented lesions have been described.3-11 These have a wide range of sensitivities and specificities, with some investigators reporting sensitivities and specificities approaching 100%. However, the diagnostic performance of a system depends on the difficulty of lesions included for analysis (measured by the median Breslow thickness of the melanoma set and the proportion of atypical nevi in the benign set) and its performance on an independent test set. Clearly, the only way to accurately compare the diagnostic performance of different clinical groups with that of SolarScan was to use the diagnostic performance of different clinical groups with that of SolarScan.

**COMMENT**

Numerous systems that automatically diagnose pigmented lesions have been described.3-11 These have a wide range of sensitivities and specificities, with some investigators reporting sensitivities and specificities approaching 100%. However, the diagnostic performance of a system depends on the difficulty of lesions included for analysis (measured by the median Breslow thickness of the melanoma set and the proportion of atypical nevi in the benign set) and its performance on an independent test set. Clearly, the only way to accurately compare the diagnostic performance of different clinical groups with that of SolarScan was to use the diagnostic performance of different clinical groups with that of SolarScan.

--

**Table 5. Correctly Classified vs Misclassified Melanomas**

<table>
<thead>
<tr>
<th>Melanoma Breslow Range, mm</th>
<th>Algorithm Diagnosis*</th>
<th>Algorithm Median Melanoma Probability, %†</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Melanoma (True Positive)</td>
<td>Benign (False Negative)</td>
</tr>
<tr>
<td>Training data set in situ</td>
<td>79</td>
<td>18</td>
</tr>
<tr>
<td>1</td>
<td>108</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>233</td>
<td>27</td>
</tr>
<tr>
<td>Test data set in situ</td>
<td>42</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>55</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>111</td>
<td>11</td>
</tr>
</tbody>
</table>

*The correctly classified lesions using an arbitrary algorithm index cutoff developed in the training set.
†The median algorithm probability of melanoma as a function of melanoma thickness category (see the “Methods” section). In the training set, a significant difference existed in the median algorithm probability of melanoma between in situ lesions, invasive melanomas thinner than 1 mm, and lesions at least 1-mm thick (P<.001; Kruskal-Wallis test). This difference failed to reach significance in the test set (P=.13).

**Table 6. Diagnostic Performance of Humans Compared With That of SolarScan**

| Group                        | Sensitivity† (Excision)‡ | Specificity§ (Nonexcision)|| | PPV (Excision) | NPV (Nonexcision) |
|------------------------------|--------------------------|-----------------------------|---|----------------|-----------------|
| Dermoscopy experts (n = 3)   | 90                       | 90                          | 59 | 45             | 32              | 26              | 96              | 95              |
| Dermatologists (n = 4)       | 81                       | 79                          | 60 | 45             | 32              | 26              | 96              | 95              |
| Trainee dermatologists (n = 3) | 85                       | 87                          | 36 (P = .006) | 34 (P = .01) | 21 (P = .005)   | 21 (P = .005)   | 92              | 93              |
| General practitioners (n = 3) | 62                       | 62                          | 63 | 61             | 26              | 25              | 89 (P = .049)   | 89 (P = .03)   |
| SolarScan                    | 85                       | 85                          | 65 | 65             | 32              | 32              | 96              | 96              |

Abbreviations: NPV, negative predictive value [probability that the patient will not have melanoma when restricted to all patients who test negative; NPV = true negative/(true negative + false negative)]; PPV, positive predictive value [probability that the patient has melanoma when restricted to those patients who test positive; PPV = true positive/(true positive + false positive)].
*Polartechnics Ltd, Sydney, Australia. High-resolution clinical and dermoscopy images and a detailed patient history from 13 melanomas and 65 benign melanocytic lesions from the independent test set (see the “Methods” section and Table 2) were used to compare the diagnostic performance of different clinical groups with that of SolarScan.
†The average sensitivity (correctly diagnosed melanomas) as a percentage.
‡The average sensitivity based on the decision to excise as correctly classifying the lesion as melanoma.
§The average specificity (correctly diagnosed nonmelanomas) as a percentage.
||The average specificity based on the decision not to excise as correctly classifying the lesion as nonmelanoma.
¶The 1-sample t test was used to compare the average sensitivity, specificity, PPV, and NPV for the diagnosis of melanoma of each clinical group with the sensitivity, specificity, PPV, and NPV determined by SolarScan.

(the probability that the lesion is melanoma when diagnosed as melanoma) and negative predictive value (the probability that the lesion is benign when diagnosed as benign) were compared with the SolarScan and clinician groups. The SolarScan positive predictive and negative predictive values were equal or superior to all clinical groups whether based on diagnosis or the decision to excise. This reached statistical significance only for the positive predictive value of trainee dermatologists and negative predictive value for general practitioners (Table 6).

For analysis of the probability that a lesion is melanoma in the melanoma set, a significantly increased average confidence (probability) of melanoma existed in all clinical groups compared with SolarScan (P<.001) (Table 7). Conversely, a significantly increased confidence (decreased probability of melanoma) by SolarScan existed compared with all clinical groups on analysis of the benign melanocytic set (P<.001).
diagnostic accuracy of systems is by directly comparing their results for the same set of lesions.

The data for algorithm development and testing were collected from 9 centers in 3 continents. Such a design increases the generalizability of the instruments’ performance. In all but 1 clinic site, the lesions collected were excised or monitored because of clinical suspicion. To reduce verification bias, lesions thought to be benign by the clinician but that required short-term digital monitoring for confirmation of their benign nature were included.3 Furthermore, a small sample of lesions were included that were clearly benign and diagnosed by classic dermoscopy features. This again reduces verification bias.

All image analysis features isolated by SolarScan were automated (ie, without input from the clinician). The clinical history features taken were modeled but not used in the final diagnostic algorithm. The only exception to the complete automated nature of the algorithm was the creation of the lesion border. Here, a 3-tiered system is used. First, an automated best-guess lesion boundary is created. If this boundary is rejected by the clinician, then a second series of automated boundaries are created. If neither of these are considered accurate, then a manual border is created by the clinician. We believe that it is an essential responsibility of the clinician to define the true lesion boundary for analysis. It is also important that the lesion border does not oversegment the lesion, that is, normal skin should not be included within the lesion boundary. If this occurs, significant differences occur with the algorithm output. For this reason, 24% of the lesions require a manual procedure to create the border. However, our results showed no significant difference in algorithm performance when comparing manual and automated boundaries.

The first diagnostic model attempted to correctly classify all pigmented benign lesions that were not melanomas. However, nonmelanocytic pigmented lesions such as seborrheic keratoses and hemangiomas were poorly discriminated. Because these lesions were weighted relatively highly in the benign set to be correctly classified during the algorithm development, it is likely that they are morphologically too similar to melanomas when using the image analysis features selected. A less important but possible contributing reason for the poor discrimination of the nonmelanocytic lesions was their relatively small sample size.

For these reasons, a model designed to discriminate only pigmented melanocytic lesions from melanoma was developed. The model was highly reproducible in the test set and gave a sensitivity of 91% and specificity of 68% for melanoma. The median probability of the melanomas in the test set was 29% and only 1.5% in the nonmelanoma set, which indicates a good separation of the 2 classes. However, because the nonmelanoma set included predominately suspicious lesions that required either excision or short-term digital monitoring for management, the true specificity in the field will be much greater. Furthermore, the median Breslow thickness was only 0.36 mm, which indicates a relatively difficult set of thin melanomas.

There is clearly a clinical limitation for an instrument that does not diagnose pigmented nonmelanocytic lesions. However, because there are strict dermoscopy criteria for distinguishing melanocytic from nonmelanocytic lesions, this clinical limitation should have less impact in a specialist setting. Nevertheless, it remains to be seen whether this is a significant limitation in general practice.

The final requirement of an automated diagnostic system is to compare its performance with diagnoses performed by humans. Although palpation of the lesion is not included for assessment by the participating clinicians, this experimental approach allows direct comparison of performance within the various clinician groups examined. It is important that none of these clinicians were involved in data collection for SolarScan algorithm development. SolarScan’s sensitivity was comparable with that of dermatoscopy experts, dermatologists, and trainee dermatologists, and had a substantially superior sensitivity (which did not, however, reach statistical significance) compared with that of general practitioners. In analysis of specificity, SolarScan’s performance was superior to that of trainee dermatologists and had a higher specificity than all 4 clinical groups (based either on absolute diagnosis or the decision to not excise a benign lesion).

The analysis of the human performance compared with that of SolarScan is somewhat limited by the relatively small sample size examined. The next stage in assessment of diagnosis by humans compared with that of SolarScan should be a formal clinical trial that incorporates both suspicious lesions and randomly selected basal lesions. Nevertheless, it seems clear from the data reported herein that SolarScan can be expected to perform well against all clinician groups in such a setting and hence would be a valuable asset for both dermatologists and primary care physicians.

The aim of this project is to produce an instrument that gives an automated diagnosis of melanoma. Because such instrumentation will never achieve 100% diagnostic accuracy, and because the gold standard of histopathologic diagnosis suffers from significant interobserver discordance, the absolute computer diagnosis will likely never be used as an absolute clinical diagnosis. Rather, it is more likely to be used as an expert second opinion, an auxiliary for clinical decision making.

Table 7. Diagnostic Performance of Humans vs SolarScan*

<table>
<thead>
<tr>
<th>Group</th>
<th>Probability of Melanoma, % (SEM)†</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Melanoma Data Set</td>
</tr>
<tr>
<td>Dermoscopy experts (n = 3)</td>
<td>80 (6.3)</td>
</tr>
<tr>
<td>Dermatologists (n = 4)</td>
<td>73 (7.8)</td>
</tr>
<tr>
<td>Trainee dermatologists (n = 3)</td>
<td>77 (7.3)</td>
</tr>
<tr>
<td>General practitioners (n = 3)</td>
<td>59 (5.5)</td>
</tr>
<tr>
<td>SolarScan</td>
<td>30 (8.5)</td>
</tr>
</tbody>
</table>

*Polartechnics Ltd, Sydney, Australia. There was a significantly increased mean probability of melanoma estimated by all clinician groups compared with SolarScan in the melanoma set (Table 2) using 1-sample t test. Conversely, there was a significant reduction in mean probability of melanoma in the benign melanocytic lesion group diagnosed with SolarScan compared with all clinician groups.

†P<.001 for all.
Accepted for Publication: May 18, 2005.

**Author Affiliations:** Sydney Melanoma Diagostic Centre, Sydney Cancer Centre, Royal Prince Alfred Hospital, Camperdown, and Faculty of Medicine (Drs Menzies and McCarthy and Ms Avramidis) and George Institute for International Health (Dr Lo), University of Sydney, Sydney, Australia; Commonwealth Scientific and Industrial Research Organisation, Mathematical and Information Sciences, Macquarie University, North Ryde, Australia (Drs Bischof and Talbot); Polartechnics Ltd, Sydney (Drs Gutenev, Mackellar, and Skladnev and Ms Wong); Victorian Melanoma Service and Department of Medicine, Alfred Hospital, Monash University, Victoria, Australia (Dr Kelly); Central Coast Skin Cancer Clinic, Toukley, Australia (Dr Cranney); Chatswood Skin Cancer Clinic, Chatswood, Australia (Dr Lye); Skin and Cancer Associates, Plantation, Fla (Dr Rabinovitz and Ms Oliviero); Skin and Cancer Foundation, Darlinghurst, Australia (Dr Virol); South East Dermatology, Carina Heights, Australia (Dr De’Ambrosis); Melanoma Unit, Princess Alexandra Hospital, Woolloongabba, Australia (Dr McCleod); Department of Dermatology, Shinshu University, Matsumoto, Nagano, Japan (Dr Koga); Department of Dermatology, University of Connecticut Health Center, Farmington (Dr Grin); Department of Dermatology, University Hospital Geneva, Geneva, Switzerland (Dr Braun); and Pigmented Lesion Clinic, School of Medicine, University of Miami, Boca Raton, Fla (Dr Johr). Dr Blum is in private practice in Konstanz, Germany.

**Correspondence:** Scott W. Menzies, MB, BS, PhD, Sydney Melanoma Diagnostic Centre, Sydney Cancer Centre, Royal Prince Alfred Hospital, Camperdown 2050, New South Wales, Australia (scott.menzies@email.cs.nsw.gov.au).

**Author Contributions:** Study concept and design: Menzies, Bischof, Talbot, Gutenev, Mackellar, and Skladnev. Acquisition of data: Menzies, Gutenev, Avramidis, McCarthy, Kelly, Cranney, Lye, Rabinovitz, Oliviero, Blum, Virol, De’Ambrosis, McCleod, Koga, Grin, Braun, and Johr. Analysis and interpretation of data: Mackellar, Lo, Gutenev, Wong, and Menzies. Drafting of the manuscript: Menzies, Gutenev, and Mackellar. Critical revision of the manuscript for important intellectual content: All authors. Statistical analysis: Wong, Lo, Mackellar, and Menzies. Obtained funding: Skladnev. Administrative, technical, and material support: Skladnev and Menzies. Study supervision: Menzies.

**Financial Disclosure:** Dr Menzies is a paid consultant for Polartechnics Ltd, the company with full ownership of the intellectual property for SolarScan. Polartechnics Ltd has filed for patents for the System and Method for Examining, Recording, and Analyzing Dermatological Conditions (US Patent filing No. 09/473270), the Boundary Finding in Dermatological Examination (US Patent filing No. 10/478077), and the Diagnostic Feature Extraction in Dermatological Examination (US Patent filing No. 10/478078). Polartechnics Ltd has filed for patients on the Diagnostic Feature Extraction in Dermatological Examination (Australian Patent application No. 20022308395 and Australian Patent No. 2003905998).

**Funding/Support:** This research was funded in part by an Australian Federal Government Research and Development Syndication Grant (13812/18/01) in 1994 and Research and Development Start Grant (STG 00186) in 1997.

**Previous Presentation:** An interim analysis of SolarScan performance (not the final data as shown herein) was presented at the American Academy of Dermatology 62nd Annual Meeting, February 2004; Washington, DC.

**REFERENCES**