Risks and Benefits of Sequential Imaging of Melanocytic Skin Lesions in Patients With Multiple Atypical Nevi

Harald Kittler, MD; Michael Binder, MD

Objective: To evaluate the utility of sequential imaging of melanocytic skin lesions.

Design: With the use of a computerized test environment, digital images of 80 melanocytic skin lesions (including 10 early melanomas) were presented to 24 dermatologists with different levels of experience in 3 sessions. The 3 sessions were designed to simulate the decision-making process (1) without the possibility of follow-up, (2) with the possibility of follow-up, and (3) after presentation of follow-up images.

Main Outcome Measures: Diagnostic performance in terms of sensitivity, specificity, accuracy, treatment threshold, and utility.

Results: The possibility of follow-up increased the treatment threshold in all groups of dermatologists compared with decision making without the possibility of follow-up. The increase of the treatment threshold was accompanied by a loss of sensitivity and a gain in specificity. The overall diagnostic accuracy remained unchanged. After presentation of follow-up images, the diagnostic accuracy improved significantly. The sensitivity improved for all readers, but the specificity improved only for the most experienced readers. The utility of sequential imaging depended on the compliance of patients with follow-up. Under the assumption that all patients are compliant with follow-up, the utility of sequential imaging was superior to decision making without follow-up over a broad range of benefit-risk ratios.

Conclusions: Sequential imaging of melanocytic skin lesions is a useful procedure for patients with multiple atypical nevi. Uncritical use of sequential imaging cannot be recommended, because the utility of this technique depends on the experience in the interpretation of follow-up images and on the patient's compliance with follow-up.

Arch Dermatol. 2001;137:1590-1595

Patients with multiple atypical nevi may have many, sometimes hundreds, of clinically abnormal-appearing nevi. The differentiation of atypical nevi from early melanoma is not always possible with certainty on clinical grounds and is complicated by the multiplicity of atypical-appearing melanocytic skin lesions in these patients. These patients also have an increased risk of developing melanoma.1-5 Therefore, the treatment of patients with multiple atypical nevi is difficult. Removal of all unusual-appearing nevi in these patients is usually not recommended, because it is impractical, involves unnecessary surgery, and does not relieve the patient from further regular skin examinations. It is generally agreed that patients with multiple atypical nevi should have regular skin examinations several times a year.3,4

Rhodes10 proposed that baseline photography and consecutive removal of suspicious new or changing nevi could reduce cost and increase reassurance in high-risk individuals, compared with wholesale surgical excision of all nevi in this group of patients.

See also pages 1641, 1575, and 1583

Recently, the use of digital photography has increased dramatically, and digital imaging offers some advantages over conventional photography.11-16 Digital epiluminescence microscopy (DELM) enhances the method with the advantages of computer technology. Images of melanocytic skin lesions are acquired and visualized electronically. Computer software supports storage, retrieval, and comparison in a time-efficient manner. Furthermore, one of the promises of this technique is to identify structural modifications by comparison over time to identify impending or incipient malignancy.11,17-19
METHODS

READERS

The 24 dermatologists participating in this study were divided into 3 groups according to their experience with epiluminescence microscopy. Group 1 (n=9) had only basic experience with epiluminescence microscopy without formal training. Group 2 (n=10) was pretrained with regard to epiluminescence microscopy but had only basic experience with DELM. Group 3 (n=5) consisted of experienced dermatologists who were pretrained in epiluminescence microscopy and routinely used DELM for the follow-up of melanocytic skin lesions.

DELM IMAGES

The DELM images were retrieved from a database currently including more than 35,000 digital images of pigmented skin lesions. The test sample consisted of 80 melanocytic skin lesions from patients with multiple atypical nevi, including 10 early melanomas. Nine melanomas lacked melanoma-specific criteria at the initial visit and were initially diagnosed as atypical nevi but showed morphologic changes over time and were therefore excised during follow-up. One melanoma was not excised at the patient’s first visit because the patient refused surgical excision. After excision, this superficial spreading melanoma measured 1.25 mm according to the Breslow method (Clark level III). All others were diagnosed as superficial spreading melanomas smaller than 0.75 mm, and 5 melanomas were in situ.

The 70 benign melanocytic skin lesions included in the study were taken at random from the 10 patients with melanoma and from 10 other randomly selected patients with multiple atypical nevi. The lesions had to meet all of the following 3 criteria: (1) availability of high-quality DELM images, (2) availability of follow-up images, and (3) histopathologic evaluation or at least 2 years of follow-up without morphologic changes during multiple examinations. Twenty lesions, including the 10 melanomas, were evaluated by excision and histopathological examination. Sixty lesions were not excised because these lesions did not show melanoma-specific criteria and did not show morphologic changes for at least 2 years of follow-up. Therefore, these 60 lesions were considered benign melanocytic skin lesions.

All DELM images were acquired with a DELM acquisition system (MoleMax II; Derma Instruments, Vienna, Austria). The pixel resolution for each digital image was 752 × 582 at 24-bit color depth.

PRESENTATION OF IMAGES

Tests were performed with a personal computer equipped with a high-quality monitor (17-in Trinitron; Sony, Tokyo, Japan). All images were presented in 24-bit color depth; the screen resolution was set to 800 × 600 pixels. The final magnification factor was 30-fold. A self-written test-assessment computer program registered the identification of the user, provided the presentation of images in random order, and recorded the individual responses of the subjects tested. The testing procedure consisted of 3 sessions. In every session the readers were asked to initiate an intervention.

In the first and second sessions, the baseline images were presented in random order. In the first session, the dermatologists had to choose between the following 2 possibilities: no intervention or excision. In the second session, the dermatologists had to choose between 3 possibilities: no intervention, follow-up with DELM, or excision.

If the readers decided to perform follow-up with DELM in the second session, the respective images were presented again, side by side with the corresponding follow-up image in the third session.

Conventional decision making without the possibility of follow-up was simulated in session 1 of the test. Sessions 2 and 3 were designed to simulate the 2 steps of decision making with the possibility of follow-up. Compared with session 1, session 2 adds an additional treatment option without changing the information provided to the readers. Session 3 simulates the additional use of follow-up information. The testing procedure is shown in Figure 1.

STATISTICAL ANALYSIS

Sensitivity, specificity, and diagnostic accuracy were calculated for each session. Sensitivity was calculated by dividing the number of excised melanomas by the total number of melanomas in the sample. Specificity was calculated as the number of nevi that were not excised divided by the total number of nevi in the sample. The diagnostic accuracy was calculated by dividing the sum of sensitivity and specificity by 2. Sensitivity, specificity, and diagnostic accuracy of session 3 were calculated by combining the responses from session 2 (for lesions that were not selected for follow-up) and session 3 (for lesions that were selected for follow-up).

The treatment threshold (T) was numerically expressed as the difference between specificity and sensitivity:

\[ T = \text{Specificity} - \text{Sensitivity} \]

T ranges from -1 to +1. T is positive if the specificity is higher than the sensitivity. T is zero if the sensitivity equals specificity and is negative if the sensitivity is higher than the specificity. In other words, as T becomes larger, the reader uses a more stringent treatment threshold (the reader increases the level of suspicion needed to initiate treatment) and vice versa.

The diagnostic utilities of 2 tests were compared by calculating the differences in their expected utilities (DU):

\[ DU = B(P)(\text{Sensitivity}_2 - \text{Sensitivity}_1) + [R(1-P)(\text{Specificity}_2 - \text{Specificity}_1)], \]

where B indicates the benefit of treatment, R indicates the risk of treatment, and P indicates the prevalence of the disease. If DU equals zero, the 2 tests perform equally well. A positive sign is in favor of test 2, and a negative sign favors test 1. We calculated the differences in the utilities of session 1 vs session 2 and session 1 vs session 3 for each reader. The treatment thresholds were fixed at the observed thresholds from each reader. We assumed that withholding treatment to a patient with melanoma is much worse than treating a patient without melanoma and therefore used benefit-risk ratios ranging from 20:1 to 200:1.

Continued on next page
The aim of this study was to compare the utility of sequential imaging of melanocytic skin lesions in patients with multiple atypical nevi by means of DELM with standard decision making without follow-up.

RESULTS

DIAGNOSTIC ACCURACY

Session 1 vs Session 2

There was no difference in the pooled diagnostic accuracy between session 1 and session 2 (0.60±0.09 vs 0.59±0.09; P=.95). Although the overall diagnostic accuracy remained unchanged, sensitivity and specificity changed substantially (Figure 2). The pooled sensitivity for session 2 was 0.45 (SD, 0.28) and significantly lower than that for session 1 (0.59±0.26; P=.02). The pooled specificity was 0.73 (SD, 0.20) for session 2 and significantly higher than that for session 1 (0.61±0.18; P<.001).

We constructed an SROC curve for session 1 and session 2 (Figure 3). Compared with the pairs of values for sensitivity and specificity observed in session 1, the pairs of values for session 2 are preferably found at the left lower region of the SROC curve, a region with high specificity and low sensitivity.

Session 1 vs Session 3

Comparing session 1 and session 3, the pooled diagnostic accuracy improved significantly (0.59±0.09 vs 0.66±0.11; P=.005), an improvement of 13.5% (95% confidence interval, 4.7%-22.4%). The improvement of diagnostic accuracy was more pronounced in the most experienced group (group 3), with an observed improvement of 22.4%, compared with 11.3% in the group with medium experience (group 2) and 11.0% in the least experienced group (group 1). The differences between groups were statistically not significant (P=.23).

The pooled sensitivity increased from 0.58 (SD, 0.23) in session 1 to 0.71 (SD, 0.20) in session 3 (P=.007). The increase in sensitivity was similar among the 3 groups of dermatologists (P for difference between groups=.83).

The pooled specificity was 0.61 (SD, 0.17) in session 1 and 0.62 (SD, 0.19) in session 3, and was not significantly different (P=.60). Groups 1 and 2 showed no improvement of specificity after presentation of follow-up images. In contrast, there was a significant gain in specificity for group 3, from 0.60 (SD, 0.09) in session 1 to 0.74 (SD, 0.18) in session 3 (P=.03).

The gain of diagnostic accuracy in session 3 compared with session 1 is visualized by constructing separate SROC curves for sessions 1 and 3 as depicted in Figure 4.

TREATMENT THRESHOLD

The pooled value for the treatment threshold (T) increased significantly from session 1 to session 2 (0.02±0.40 vs 0.29±0.46; P<.001). As shown in Figure 5, the increase was more pronounced in group 3, although the differences between groups were statistically not significant (P=.12). After presentation of follow-up images in session 3, the treatment thresholds were not significantly different from the values observed in session 1 (P=.11).

UTILITY

Assuming a benefit-risk ratio of 20:1, we observed a significant loss of utility for session 2 compared with session 1 (DU, −0.2; 95% confidence interval, −0.01 to −0.48; P=.04) and a significant gain in utility for session 3 compared with session 1 (DU, 0.36; 95% confidence interval, 0.12-0.60; P=.005). As shown in Figure 6, increasing the benefit-risk ratio increased the gain in utility for session 3 and simultaneously increased the loss in utility for session 2.
With the use of a computerized test environment, 24 dermatologists of varying skills were tested on 80 images of melanocytic skin lesions that were documented with DELM. All images were clinically atypical nevi or early melanomas drawn from 20 patients with multiple atypical nevi. All readers were informed that only atypical nevi or melanomas would be presented during the test. The goal was to identify true melanomas (Figure 7). The 3 sessions were designed to simulate the decision-making process of dermatologists in the following situations: (1) without the possibility of follow-up, (2) with the possibility of follow-up, and (3) after presentation of follow-up images.

With regard to diagnostic accuracy, there was no difference between session 1 and session 2. This was expected, because session 2 only added an additional treatment option without changing the amount of information. Although the diagnostic accuracy did not differ between session 1 and session 2, we observed a significant decrease in sensitivity and a significant increase in specificity. The reason for this is that the dermatologists increased their treatment (excision) thresholds in session 2 compared with session 1. In other words, with the possibility of follow-up, dermatologists increased the level of suspicion needed to excise a lesion. Interestingly, this effect was more pronounced among readers who were experienced in the field of DELM.

We also provided evidence that, in comparison with decision making without the possibility of follow-up, the possibility of follow-up significantly increased the diag-

Figure 2. Mean values for sensitivity (A) and specificity (B) of 24 readers according to the test session. Note the decrease in sensitivity in session 2 compared with session 1, which is mirrored by an increase in specificity. Error bars indicate SD.

Figure 3. Mean values for the treatment threshold T (specificity–sensitivity) by test session. The zero line indicates equal values for sensitivity and specificity. Compared with session 1, T increases in session 2 in all groups of dermatologists, which is equivalent to an increase of the treatment threshold. This effect is most pronounced in group 3. Group 1 has the least experience; group 2, medium experience; and group 3, most experience. Error bars indicate SD.

Figure 4. Summary receiver operating characteristic curve for session 1 and session 2. Note that the pairs of values for sensitivity and specificity from session 2 tend to lay at the left lower region of the curve, a region with high specificity and low sensitivity.

Figure 5. Separate summary receiver operating characteristic curves for session 1 and session 3. Compared with session 1, the curve from session 3 bends up to the desirable left upper part of the receiver operating characteristic space.
potential clinical implications for therapeutic decisions. Therefore, in our study, calculation of the diagnostic accuracy was based on the therapeutic interventions chosen by the readers. A diagnostic test is clinically relevant if it contributes to a correct therapeutic decision, taking into account the benefit and risks of the available treatment. To initiate treatment, the reader must compare in an intuitive way the benefit of treating patients with melanoma with the risks of treating patients without melanoma. The reader will choose a treatment threshold according to the intuitive comparison of this ratio. If the net benefit of treatment is higher than the net risk of treatment, the reader will choose a treatment threshold with higher sensitivity and lower specificity and vice versa. Because sensitivity and specificity are in constant tension, increasing 1 of the 2 will decrease the other. This is equivalent to choosing different operating points on an SROC curve. The trade-off between sensitivity and specificity (the treatment threshold) and the benefit-risk ratio of treatment will have important implications on the utility of a diagnostic test. The utility is a more relevant measure of the clinical performance of a diagnostic test, because it takes into account the prevalence of the disease, the diagnostic accuracy, the treatment threshold, and the benefit-risk ratio of treatment.

An important finding of our study is that the utility of sequential imaging will depend on whether the patients are compliant with the follow-up regimen. Session 2 and session 3 simulate 2 extreme situations. Session 2 is equivalent to decision making with the possibility of follow-up when all patients who were selected for follow-up are unavailable for follow-up. Session 3 is different from session 2 in that it simulates that all patients who were selected for follow-up are compliant with the follow-up regimen. We showed that there is a significant gain in utility for session 3 compared with session 1 and a significant loss in utility for session 2 compared with session 1. We can think of this as a form of upper and lower limits for the utility of the follow-up procedure. Regardless of the skills of the dermatologists, if all patients are unavailable for follow-up, the utility of sequential imaging will be worse than the utility of decision making without the possibility of follow-up. On the other hand, if all patients are compliant with follow-up, the utility of sequential imaging is superior to standard decision making without follow-up.

We also examined the impact of changes in the benefit-risk ratio on the differences in the utility. By doing this, we demonstrated that increasing the benefit-risk ratio increased the gain in utility for session 3 and simultaneously increased the loss in utility for session 2.

For most clinical situations, it is important to evaluate the values of diagnostic tests in relation to their potential clinical implications for therapeutic decisions. Therefore, in our study, calculation of the diagnostic accuracy was based on the therapeutic interventions chosen by the readers. A diagnostic test is clinically relevant if it contributes to a correct therapeutic decision, taking into account the benefit and risks of the available treatment. To initiate treatment, the reader must compare in an intuitive way the benefit of treating patients with melanoma with the risks of treating patients without melanoma. The reader will choose a treatment threshold according to the intuitive comparison of this ratio. If the net benefit of treatment is higher than the net risk of treatment, the reader will choose a treatment threshold with higher sensitivity and lower specificity and vice versa. Because sensitivity and specificity are in constant tension, increasing 1 of the 2 will decrease the other. This is equivalent to choosing different operating points on an SROC curve. The trade-off between sensitivity and specificity (the treatment threshold) and the benefit-risk ratio of treatment will have important implications on the utility of a diagnostic test. The utility is a more relevant measure of the clinical performance of a diagnostic test, because it takes into account the prevalence of the disease, the diagnostic accuracy, the treatment threshold, and the benefit-risk ratio of treatment.

An important finding of our study is that the utility of sequential imaging will depend on whether the patients are compliant with the follow-up regimen. Session 2 and session 3 simulate 2 extreme situations. Session 2 is equivalent to decision making with the possibility of follow-up when all patients who were selected for follow-up are unavailable for follow-up. Session 3 is different from session 2 in that it simulates that all patients who were selected for follow-up are compliant with the follow-up regimen. We showed that there is a significant gain in utility for session 3 compared with session 1 and a significant loss in utility for session 2 compared with session 1. We can think of this as a form of upper and lower limits for the utility of the follow-up procedure. Regardless of the skills of the dermatologists, if all patients are unavailable for follow-up, the utility of sequential imaging will be worse than the utility of decision making without the possibility of follow-up. On the other hand, if all patients are compliant with follow-up, the utility of sequential imaging is superior to standard decision making without follow-up.

We also examined the impact of changes in the benefit-risk ratio on the differences in the utility. By doing this, we demonstrated that increasing the benefit-risk ratio increased the gain in utility for session 3 and simultaneously increased the loss in utility for session 2.

A possible limitation of our study is the experimental design using computer simulation and the potential divergence from real-world situations. Under real-world conditions, dermatologists may behave differently with regard to their therapeutic decisions. We are convinced that these possible limitations are outweighed by the advantages provided by the standardized and reproducible setting. The experimental design was also helpful in minimizing the influence of confounding variables. With regard to the diagnostic accuracy, the observed values may not be comparable to val-

Figure 6. Differences in the expected utility (DU) between session 1 and session 2 and between session 1 and session 3 at increasing benefit-risk ratios. The zero line indicates no differences between the test sessions (dotted line). Positive values for DU indicate a gain in utility and negative values indicate a loss in utility. Increasing the benefit-risk ratio increased the gain in utility for session 3 compared with session 1 and simultaneously increased the loss in utility for session 2 compared with session 1.

Figure 7. Sequential digital epiluminescence microscopic images of an early melanoma (superficial spreading melanoma, 0.3-mm Breslow thickness). The left image (A) was recorded 5 months before the right image (B). The lesion shows an asymmetric increase in size as well as structural changes. The magnifications of images A and B are identical; the bar indicates 1 mm.

For most clinical situations, it is important to evaluate the values of diagnostic tests in relation to their potential clinical implications for therapeutic decisions. Therefore, in our study, calculation of the diagnostic accuracy was based on the therapeutic interventions chosen by the readers. A diagnostic test is clinically relevant if it contributes to a correct therapeutic decision, taking into account the benefit and risks of the available treatment. To initiate treatment, the reader must compare in an intuitive way the benefit of treating patients with melanoma with the risks of treating patients without melanoma. The reader will choose a treatment threshold according to the intuitive comparison of this ratio. If the net benefit of treatment is higher than the net risk of treatment, the reader will choose a treatment threshold with higher sensitivity and lower specificity and vice versa. Because sensitivity and specificity are in constant tension, increasing 1 of the 2 will decrease the other. This is equivalent to choosing different operating points on an SROC curve. The trade-off between sensitivity and specificity (the treatment threshold) and the benefit-risk ratio of treatment will have important implications on the utility of a diagnostic test. The utility is a more relevant measure of the clinical performance of a diagnostic test, because it takes into account the prevalence of the disease, the diagnostic accuracy, the treatment threshold, and the benefit-risk ratio of treatment.

An important finding of our study is that the utility of sequential imaging will depend on whether the patients are compliant with the follow-up regimen. Session 2 and session 3 simulate 2 extreme situations. Session 2 is equivalent to decision making with the possibility of follow-up when all patients who were selected for follow-up are unavailable for follow-up. Session 3 is different from session 2 in that it simulates that all patients who were selected for follow-up are compliant with the follow-up regimen. We showed that there is a significant gain in utility for session 3 compared with session 1 and a significant loss in utility for session 2 compared with session 1. We can think of this as a form of upper and lower limits for the utility of the follow-up procedure. Regardless of the skills of the dermatologists, if all patients are unavailable for follow-up, the utility of sequential imaging will be worse than the utility of decision making without the possibility of follow-up. On the other hand, if all patients are compliant with follow-up, the utility of sequential imaging is superior to standard decision making without follow-up.

We also examined the impact of changes in the benefit-risk ratio on the differences in the utility. By doing this, we demonstrated that increasing the benefit-risk ratio increased the gain in utility for session 3 and simultaneously increased the loss in utility for session 2.

A possible limitation of our study is the experimental design using computer simulation and the potential divergence from real-world situations. Under real-world conditions, dermatologists may behave differently with regard to their therapeutic decisions. We are convinced that these possible limitations are outweighed by the advantages provided by the standardized and reproducible setting. The experimental design was also helpful in minimizing the influence of confounding variables. With regard to the diagnostic accuracy, the observed values may not be comparable to val-
ues observed in real clinical situations. However, our main interests focused on the differences between the test sessions and not on the absolute values.

Accepted for publication April 3, 2001.
This study was supported by grant FWF-P11735MED from the Austrian Science Fund, Vienna.
Corresponding author and reprints: Harald Kittler, MD,
Department of Dermatology, University of Vienna Medical School, Waehringerguertel 18-20, A-1090 Vienna, Austria (e-mail: h.kittler@akh-wien.ac.at).

REFERENCES


News and Notes

Photomedicine Society Meeting will be held on February 21, 2002, in New Orleans, La. For information, contact Susan Milberger at fax No. (214) 648-0280 (e-mail: susanmilberger@utsouthwestern.edu).