Applying a decision support system in clinical practice: Results from melanoma diagnosis

Stephan Dreiseitl[*‡], PhD, Michael Binder[†], MD, Staal Vinterbo[•], PhD, Harald Kittler[†], MD

[*] Dept. of Software Engineering
Upper Austria University of Applied Sciences, Hagenberg, Austria

[†] Dept. of Dermatology
Medical University of Vienna, Vienna, Austria

[•] Decision Systems Group
Brigham & Women's Hospital, Harvard Medical School, Boston, MA

[†] Dept. of Biomedical Engineering
University of Health Sciences, Medical Informatics and Technology, Hall, Austria

Abstract

The work reported in this paper investigates the use of a decision-support tool for the diagnosis of pigmented skin lesions in a real-world clinical trial with 511 patients and 3827 lesion evaluations. We analyzed a number of outcomes of the trial, such as direct comparison of system performance in laboratory and clinical setting, the performance of physicians using the system compared to a control dermatologist without the system, and repeatability of system recommendations.

The results show that system performance was significantly less in the real-world setting compared to the laboratory setting (c-index of 0.87 vs. 0.94, $p = 0.01$). Dermatologists using the system achieved a combined sensitivity of 85% and combined specificity of 95%. We also show that the process of acquiring lesion images using digital dermoscopy devices needs to be standardized before sufficiently high repeatability of measurements can be assured.

Introduction

The advancement of medical knowledge is based on the timely dissemination of reviewed research findings that were obtained by following a set of quality criteria in study design, management, and evaluation. This is no different for predictive models built from medical data sets: one can only expect them to be widely used if a set of quality criteria is met by these models as well. Our own methodology review[1] found that many of the criteria which would allow a reader to judge the quality of a model (in particular, details of the model building process) are rarely reported in a satisfactory manner.

There are a number of papers in the literature that focus on model validation in medical domains.[2–4] The conclusions of these papers are mostly positive, in the sense that systems validated internally (i.e., on data from the same patient sample as had been used to build the system) are indeed beneficial when deployed clinically. The use of such systems also raises the question of how large an influence they have on physician performance. This question has also been considered in the literature.[5–8]

In this paper, we investigate the performance of a clinical decision-support system (CDSS) that had been previously built and validated internally. The domain of discourse is dermatology, in particular the diagnosis of pigmented skin lesions. The diagnostic technique of choice in this field is *dermoscopy*, also known as *dermatoscopy* or *epiluminescence microscopy*. Dermoscopy is an imaging technique that uses polarized light to make pigmented lesion structures in the epidermis and the papillary dermis more easily visible. Digital dermoscopy devices are generally bundled with computer systems that serve as digital image repositories. The advantage of this approach is the ease with which the images can be manipulated (e.g., enlarged), archived and retrieved for follow-up consultations, and the ability to include CDSS in the diagnostic process.

In numerous controlled experiments, it was shown that dermoscopy is well-suited to increase the diagnostic performance of dermatologists.[9,10] It has, however, been noted that some training and experience is required to actually reap the benefits of this technology.[11] As an added bonus, the easy availability of large data sets has led to the structured analysis of lesion images. A review of machine-learning approaches to dermoscopy is available in the literature.[12]

Artificial neural networks, support vector machines, and logistic regression are machine learning algorithms that are well suited to the problem of diagnosing pigmented skin lesions.[13] For the work reported here, we had previously built and validated a neural network model, based on 1311 lesion pictures taken at the Dept. of Dermatology of the Medical University of Vienna using a MoleMax II dermoscopy instrument (Derma Medical Systems, Vienna, Austria). Of the 1311 lesions considered, 125 were melanomas, and 1186 were benign lesions. Images taken at a resolution of 752×582 pixels and 24 bit color depth were segmented with a local thresholding algorithm. A feature extraction step resulted in a lesion description as a vector of 29 real-valued components. The discriminatory power of the trained neural network was 0.94, as measured by the area under the ROC curve.

The purpose of the study reported here is to investigate the performance of an internally validated CDSS in a clinical setting, and the performance of physicians using this CDSS as a tool for providing a second opinion on lesion malignancy. By using a study design that has two physicians using the CDSS on the same patient, we were also able to assess the repeatability of system outputs. More details on the study setup are given in the next section.

Material and methods

In the year 2004, a clinical trial using a CDSS built upon the neural network classifier described above took place at the Dept. of Dermatology of the Medical University of Vienna. This department serves as a secondary and tertiary referral center; prior probability of melanoma is thus higher than in the general population. The study design of the clinical trial was as follows: Every patient participating in the trial was examined by two study physicians (out of a group of six). The study physicians had between zero and four years of expertise in dermatology, and participated in the trial based on availability. Informed consent to participate in the trial was obtained from both the patients as well as the study physicians. The physicians used special dermoscopy equipment which was linked to a CDSS that provided a malinancy rating for each lesion examined. The CDSS output was shown on a second computer monitor in the form of a slider position on a continuous-scale malignancy rating bar. Internally, the neural network produced an output in the range of 0 to 1, with 0 corresponding to benign lesions, and 1 corresponding to malignant lesions. The malignancy rating bar was color-coded to provide a further visual impression of the output: The range 0–0.1 was colored green, 0.1–0.4 yellow, and 0.4–1 red. These ranges were chosen in an arbitrary manner.

The physicians could use or ignore the CDSS output at their own discretion. The physicians were not told the exact performance details of the system. However, they were aware from the literature that systems of this kind perform at about the the level of expert dermatologists. The physicians could also choose which lesions to examine. For every lesion, the study physicians gave a dichotomous evaluation by rating each lesion as either benign or malignant. In addition to the two examinations by the study physicians, all the study patients were also seen by an expert dermatologist who examined them with regular dermoscopic equipment. The role of this physician was that of a safe-guard, should the study physicians (who were not as experienced) miss a malignant lesion. Patient management was performed according to the recommendations of this expert dermatologist. The gold standard for lesions was determined by histopathology for all excised lesions and one-year follow-up for all lesions considered benign and thus not excised.

A consecutive sample of 511 patients participated in the study; 3827 lesion examinations were undertaken. In 786 cases, a lesion was examined by both physicians. The number of lesion examinations ranged from 1370 lesions examined on 339 patients for the most active physician to 122 lesion on 45 patients for the least active (with median values of 379.5 lesion examinations on 95.5 patients). A number of patients and lesion examinations had to be removed from the study. This happened mainly when an examination could not be attributed to a physician, or when the gold standard of a lesion examination could not be determined due to follow-up exams not being made. After removal of incomplete cases, the number of patients remaining in the study was 458, with 3021 lesion examinations. This group consisted of 431 healthy patients, and 27 patients with at least one melanoma.

The outcomes considered in this study were as follows: performance of the CDSS as a stand-alone system in a clinical environment; performance of the CDSS-assisted physicians, compared with the expert without CDSS; and repeatability of system recommendations. These outcome measures can be broken down into a number of smaller points that will be considered separately. For each of these measures, the unit of discourse is one lesion examination. This means that we focus attention on the aspect of the study that

Table 1: Sensitivities and specificities of the six study physicians using the CDSS, of the expert dermatologist without CDSS, and of the CDSS by itself. System outputs were dichotomized at a the threshold of the classifier closest to the ideal classifier.

	Physician ID								
	1	2	3	4	5	6	combined	expert	CDSS
sensitivity	78%	89%	50%	82%	100%	100%	85%	97%	75%
specificity	95%	95%	97%	94%	95%	80%	95%	93%	84%

is most directly related to the CDSS, and ignore the entity “patient” that usually consists of more that one lesion evaluation.

Results

CDSS as stand-alone device: Using the 3021 malignancy ratings in the study, we found that the neural network in the CDSS can distinguish between benign and malignant cases at a discriminatory level of 0.87 (95% confidence interval [0.82, 0.92]), as measured by the area under the ROC curve. While this value is roughly comparable to values reported for similar systems in the literature, it is significantly less than the value obtained during internal validation of the system (0.94, $p = 0.01$).

The system output has to be dichotomized for direct comparison to expert lesion assessments, which are either benign or malignant. Using the point on the ROC curve closest to the ideal classifier (i.e., closest to the upper-left corner of the unit square), we obtain sensitivity and specificity values of 75% (95% CI [62%, 85%]) and 84% (95% CI [83%, 86%]), respectively.

Comparing physicians with and without CDSS: When a physician does not even examine a lesion, it is impossible for the CDSS to issue an alert, because the system can only provide assessments for lesions that are examined using the dermoscopy equipment. The event of a study physician not even examining a melanoma occurred at least three times. The number may be even higher, but in three cases did the expert dermatologist identify melanomas that were missed by the study physicians.

A summary of the diagnostic performance of all six study physicians is given in Table 1. The physicians are sorted with respect to number of lesion examinations, with physician 1 being the most active (1370 examinations), and physician 6 being the least active (122 examinations). For comparison, the table also includes the performance of the expert dermatologist, and of the CDSS when considered as a stand-alone system (dichotomized as described above).

Table 2: Contingency table for correctness of system and physician malignancy assessments. System outputs were dichotomized at a the threshold of the classifier closest to the ideal classifier.

	phys. corr.	phys. incorr.	
CDSS corr.	2441	104	2545
CDSS incorr.	410	66	476
	2851	170	3021

Disagreement between physicians and system: The system output, as a number on a continuous scale, first needs to be dichotomized in order to be comparable to a physician assessment of benign or malignant. If we perform this dichotomization at the threshold where the corresponding classifier is closest to the upper-left corner of the unit square, we arrive at the contingency table shown in Table 2. In this table, we pooled the assessments of all six study physicians. One can observe that there is a large agreement between system and physician recommendations. In only 66 of the 3021 lesion examinations (2.2%) did both system and physician miss the correct diagnosis. Of these 66, 5 (7.6%) were melanomas.

It is also of interest to determine how often the physicians agreed with the system, stratified by system malignancy assessment. For this analysis, we used the threshold of 0.4, at which the system output slider moves from the yellow to the red region of the output bar. There were 140 lesion examinations that the system rated as malignant; in 59 of these cases, the physi-

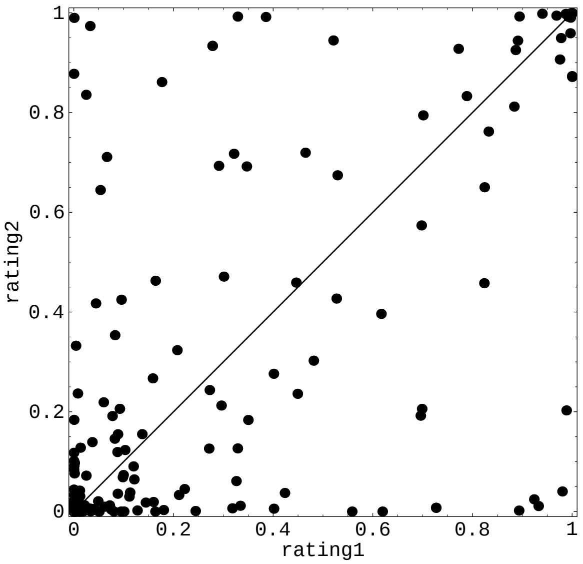

Figure 1: Scatterplot of the decision-support system output for two examinations of the same lesion, for a total of 768 lesions.

cian was of the same opinion (42%). For the 2881 cases in which the system rating was benign, the users agreed in 2734 cases (95%).

Reproducability of system outputs: By having more than one physician examine a patient, we obtained a total of 786 lesions that were examined by both physicians. Figure 1 shows a scatterplot of system outputs, with the first rating on the x-axis, and the second rating on the y-axis. Note that 698 of 786 ratings are in the lower left part of the scatterplot, the area $[0, 0.2] \times [0, 0.2]$. While the agreement is indeed visible for lesions with a benign rating, one can observe that there is considerable spread in the region where at least one of the ratings is greater than 0.2. Indeed, if we consider the region where one of the ratings is larger than 0.8, there is only an about 70% chance that the other rating is also above 0.8.

Discussion

The study presented here investigated the deployment of a decision support system in a clinical environment. Note that we did not consider the clinically relevant question of lesion management: In our study, the diagnosis of melanoma was equivalent to an excision recommendation, and vice versa.

We observed that the machine-learning component of the system, taken by itself, did not perform on the level that it achieved during internal validation. There are a number of reasons why this could be the case: A change in patient population, a change in imaging modalities, or a change in lesion characteristics. While we can rule out the first (model building and application used a convenience sample of patients presenting at the Dept. of Dermatology of the Medical University of Vienna), the other two factors may well have had a significant impact on model performance. The non-calibrated use of imaging equipment, as was performed here, was recently shown to have a significant impact on color perception and, subsequently, lesion evaluation.[14] Lesion characteristics may also have changed, in the sense that the sample used for model building was collected by experts, who may have considered lesions different from those now selected by the inexperienced study physicians who applied the CDSS.

Although the performance of the system was less than expected, we nevertheless found that a group of dermatologists with varying levels of expertise in dermoscopy, when using the system, achieved sensitivities and specificities that are on par with numbers reported in the literature.[11] Indeed, when compared with an expert dermatologist, the physicians taken together exceeded the specificity of the expert, although at the cost of decreased sensitivity. In a screening environment, however, higher sensitivity is preferable to higher specificity.

It must be emphasized that a CDSS as presented here cannot be used as an automated alert system, because it is only invoked on lesions that the physician wants to investigate. This limitation was evident in our study, where a total of three melanomas were missed by the study physicians, because they did not even investigate the lesions. The new technique of digital total body photography, when used in conjunction with a CDSS, may remedy this situation, because such a system would have access to *all* the lesion images of a patient.

Limitations: We expected a high degree of agreement between system outputs for the same lesion, because the neural network was trained on features that should not change in multiple examinations (such as asymmetry, border, color distributions, among others). It was therefore surprising that CDSS outputs of dermoscopy images are not as reproducible as desired. Changes in imaging modality are not likely to account for this, although the thermal nature of an imaging device can change rapidly, resulting in a shift in spectral power distribution and a subsequent change in color perception. Although this may be one of the reasons for the observed poor reproducibility, we believe that the very act of image acquisition is not sufficiently robust for repeatable measurements: because the image is taken by placing the dermoscopy equipment directly on the skin of the patient, there are variations in tilt and force that may be sufficient to significantly change the characteristics of the image. In visual inspection, there may be little difference between images of the same lesion visible to the human eye. This, however, may be due to the way that the eye and the human brain can compensate for changes in a way that is very hard to duplicate for a machine.

Conclusion

In this paper, we reported on the outcomes of a study investigating the application of a decision-support system to melanoma diagnosis. The main findings of the study were the fact that the CDSS by itself did not perform as well as expected. Further work will be required to standardize the image acquisition process, as evidenced by the poor reproducability of results on lesions with at least one high malignancy rating.

References

[1] Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform 2002;35:352–359.

[2] Sintchenko V, Iredell JR, Gilbert GL, Coiera E. Handheld computer-based decision support reduces patient length of stay and antibiotic prescribing in critical care. J Am Med Inform Assoc 2005;12(4):398–402.

[3] Dayton CS, Ferguson JS, Hornick DB, Peterson MW. Evaluation of an Internet-based decision-support system for applying the ATS/CDC guidelines for tuberculosis preventive therapy. Med Decis Making 2000;20(1):1–6.

[4] Chang PL, Li YC, Wang TM, Huang ST, Hsieh ML, Tsui KH. Evaluation of a decision-support system for preoperative staging of prostate cancer. Med Decis Making 1999;19(4):419–427.

[5] Dreiseitl S, Binder M. Do physicians value decision support? A look at the effect of decision support systems on physician opinion. Artif Intell Med 2005;33:25–30.

[6] Friedman C, Gatti G, Elstein A, Franz T, Murphy G, Wolf F. Are clinicians correct when they believe they are correct? Implications for medical decision support. Medinfo 2001;10(Pt 1):454–458.

[7] Berner ES, Maisiak RS. Influence of case and physician characteristics on perceptions of decision support systems. J Am Med Inform Assoc 1999;6(5):428–434.

[8] Friedman CP, Elstein AS, Wolf FM, Murphy GC, Franz TM, Heckerling PS, Fine PL, Miller TM, Abraham V. Enhancement of clinicians' diagnostic reasoning by computer-based consultation: a multisite study of 2 systems. JAMA 1999; 282(19):1851–1856.

[9] Westerhoff K, McCarthy WH, Menzies SW. Increase in the sensitivity for melanoma diagnosis by primary care physicians using skin surface microscopy. British Journal of Dermatology 2000;143:1016–1020.

[10] Carli P, Giorgi VD, Crocetti E, Mannone F, Massi D, Chiarugi A, Giannotti B. Improvement of malignant/benign ratio in excised melanocytic lesions in the 'dermoscopy era': a retrospective study 1997–2001. British Journal of Dermatology 2004;150:687–692.

[11] Kittler H, Pehamberger H, Wolff K, Binder M. Diagnostic accuracy of dermoscopy. Lancet Oncol 2002;3(3):159–165.

[12] Binder M, Kittler H, Pehamberger H, Dreiseitl S. Differentiation between benign and malignant skin tumors by image analysis, neural networks, and other methods of machine learning. In: Wilhelm KP, et al., editors, Bioengineering of the Skin: Skin Imaging and Analysis, 2nd edition. Informa Healthcare, 2006; pp. 297–304, pp. 297–304.

[13] Dreiseitl S, Ohno-Machado L, Vinterbo S, Billhardt H, Binder M. A comparison of machine learning methods for the diagnosis of pigmented skin lesions. J Biomed Inform 2001;34:28–36.

[14] Grana C, Pellacani G, Seidenari S. Practical color calibration for dermoscopy, applied to a digital epiluminescence microscope. Skin Res Technol 2005;11(4):242–247.