Visual Cues Do Not Improve Lesion ABC(D) Grading

Visual Cues Do Not Improve Lesion ABC(D) Grading

Matteo Zanotto[a,b], Lucia Ballerini[b], Ben Aldridge[c], Robert B. Fisher[b], Jonathan Rees[c]

[a]Istituto Italiano di Tecnologia, Via Morego 30, Genova, Italy;
[b]School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh, UK;
[c]Department of Dermatology, University of Edinburgh, Lauriston Place, Edinburgh, UK

ABSTRACT

In this work evidence is presented supporting the hypothesis that observers tend to evaluate very differently the same properties of given skin-lesion images. Results from previous experiments have been compared to new ones obtained where we gave additional prototypical visual cues to the users during their evaluation trials. Each property (*colour*, *colour uniformity*, *asymmetry*, *border regularity*, *roughness of texture*) had to be evaluated on a 0–10 range, with both linguistic descriptors and visual references at each end and in the middle (e.g. light/medium/dark for colour). A set of 22 images covering different clinical diagnoses has been used in the comparison with previous results. Statistical testing showed that only for a few test images the inclusion of the visual anchors reduced the variability of the grading for some of the properties. Despite such reduction, though, the average variance of each property still remains high even after the inclusion of the visual anchors. When considering each property, the average variance significantly changed for the *roughness of texture*, where the visual references caused an increase in the variability. With these results we can conclude that the variance of the answers observed in the previous experiments was not due to the lack of a standard definition of the extrema of the scale, but rather to a high variability in the way observers perceive and understand skin-lesion images.

1. INTRODUCTION

In many medical imaging domains, users' performance in the evaluation of images is critical for the success of the diagnostic process. This is particularly true in specific disciplines, like dermatology, where the use of qualitative guidelines relies heavily on people's ability to describe what they see according to some pre-defined concepts. This work is a case study of observer performance. The aim is investigating how accurately people can describe skin-lesion images according to qualitative properties suggested by guidelines like the ABC(D) rule.[1]
To do so, users' feedback has been collected through a web page (see section 3.1) specifically designed to cover a set of five properties, three of which are proposed by the ABC(D) rule (see section 2). The analysis of the collected data (see section 3.2) demonstrates how evaluations obtained through qualitative guidelines show a very high variability due to users' subjectivity in the interpretation of the assessed qualities (see section 4).

2. THE ABC(D) RULE

The ABC(D) rule was proposed as a mnemonic in 1985 by Friedman *et al.*,[1] to aid both clinicians and laypeople in the earlier diagnosis of melanomas. The mnemonic was based on criteria, which in the authors' experience, tended to be apparent in melanoma. Over the 25 years since it's inception additional features have been suggested to improve its' utility,[2] but fundamentally the following 4 property rule remains at the heart of most public educational campaigns:

- **Asymmetry** as melanomas tend to be asymmetric both in shape and in terms of colour distribution
- **Border irregularity** as melanomas have less defined and more jagged borders than benign lesions
- **Colour variegation** as melanomas tend to have a non-uniform colour distribution
- **Diameter** as melanomas tend to be wider than 6 mm

Further author information: (Send correspondence to L. Ballerini)
Lucia Ballerini: E-mail: lucia.ballerini@ed.ac.uk, Telephone: +44 (0)131 651 5664

Zanotto, M., Ballerini, L., Aldridge, B., Fisher, B., & Rees, J. (2011). Visual Cues Do Not Improve Lesion ABC(D) Grading. In Proceedings SPIE Medical Imaging Vil 7966. (pp. 796600-1 - 796600-10). doi: 10.1117/12.878115

Major stress is put on the fact that the sooner melanomas are identified, the higher the probability of effective surgical removal, which translates to a higher survival rate. For this reason, people are encouraged to actively examine their skin following the ABC(D) rule in search for suspicious signs which might suggest the development of melanoma.
Over the years studies have been conducted on the effectiveness of the ABC(D) rule, such as Brändström *et al.*,[3] Gunasti *et al.*,[4] Meyer *et al.*,[5] Reetz Müller *et al.*[6] While some papers[3,6] claim that the use of the ABC(D) rule had a positive impact on the answers given by the participants, others[4,5] point out a substantial variability in the way different people assess some of the criteria. Even the results obtained by Laskaris in his Master's thesis[7] support the claim suggesting that people evaluate the same skin lesions in different ways.
Given the importance of a correct evaluation of the four key properties in order to obtain successful results with the ABC(D) rule, the findings presented in some of the referenced works[4,5,7] highlight the necessity for a more extensive and closer study of people's assessment performance. An attempt to gain a better understanding of the variability of users' evaluation is the essence of this work and will be further detailed in the coming sections.

3. USER PERFORMANCE WITH THE ABC(D) RULE

The ABC(D) rule, presented in the previous section, relies strongly on the assumption that people can effectively describe what they see in terms that, albeit qualitative, show consistency across different observers. This is true for all the qualitative rules currently in use in dermatology either to support the diagnostic process or as self screening guidelines.
During experiments reported in Laskaris's research,[7,8] evidence emerged suggesting a substantial variability in the assessment different people give when evaluating characteristics of the same skin-lesion image. Specifically, it was observed that when asked to assess five properties of skin-lesion images (namely *colour*, *colour uniformity*, *asymmetry*, *border regularity* and *roughness of texture*) people gave very different evaluations for the same picture. The main task of the research reported here was to investigate more rigorously the consistency of the qualitative judgement people provide when presented with a skin-lesion image, in order to understand whether the variability observed previously was caused by the specific experimental set-up or rather by a real difference in the way each person interprets the concepts given by the guidelines.
In previous experiments people were asked to rank the qualities by moving a slider on a scale having only linguistic expressions as references (e.g. light/medium/dark for colour). The problem with using only a linguistic approach is that it is impossible to separate the effects of differences in users' subjective interpretation of the extrema from their intra-person intrinsic variability. In order to isolate the two, a new experiment has been performed, modifying the interface with the addition of visual anchor points to the linguistic references. A new image-set has been acquired for the experiment by the Department of Dermatology of the University of Edinburgh. While nearly half of the images were the same of those used in Laskaris's research[7] (see Figure 1), new ones have been introduced in order to get a more balanced representation of the main diagnostic classes. The new images were used to perform additional studies on the distribution of the answers which have been presented elsewhere.[9,10]
As three of the tested characteristics (*asymmetry*, *border regularity* and *colour uniformity*) are the first three elements of the ABC(D) rule described in section 2, this experiment constitutes an empirical study on users' ability to follow qualitative visual guidelines on dermatological images.

3.1 Experimental Set-up

In order to guarantee comparability with the previously obtained answers,[7] the web interface used for collecting data has been kept substantially unchanged. The page (see Figure 2) is structured to present 45 skin-lesion images to the user in a randomly selected order, requiring the assessment of the 5 previously mentioned properties: *colour*, *colour uniformity*, *asymmetry*, *border regularity* and *roughness of texture*. The evaluation is provided through the use of a set of analogue sliders (one for each property) which can be moved left-to-right producing an associated score in the continuous 0–10 range. In order to correctly understand the numerical values presented in the following analysis it is useful to remember that the endpoints for each property are:

Figure 1. Skin lesion images on which the analysis has been performed.

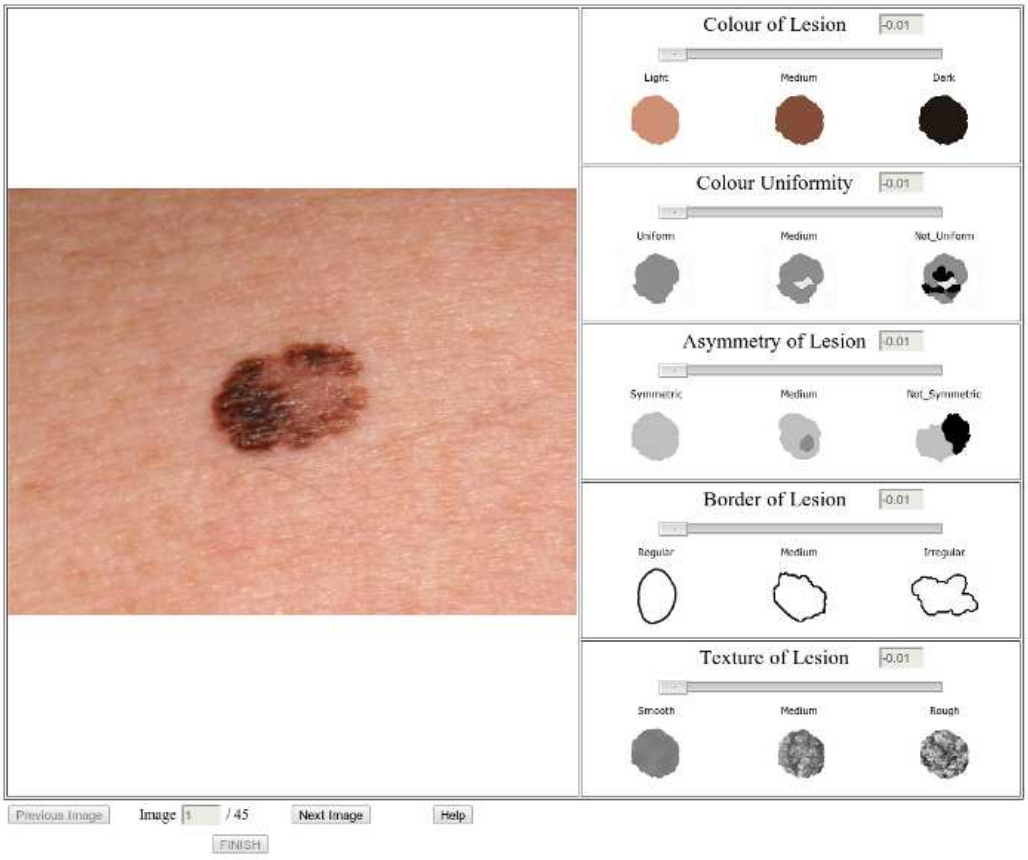

Figure 2. Web interface used for data collection.

colour	0 = light	10 = dark
colour uniformity	0 = uniform	10 = not uniform
asymmetry	0 = symmetric	10 = asymmetric
border regularity	0 = regular	10 = irregular
roughness of texture	0 = smooth	10 = rough

The random ordering was introduced to minimise any evaluation bias due to specific image sequences. The bias is mainly due to the fact that people tend to adjust their evaluations on the basis of what they have previously seen, often considering, in the assessment, the evaluation given to previous samples. The effect is greater in cases where people are presented with subjects they are not familiar with, as the aid provided by prior knowledge is limited. While the randomisation cannot eliminate the bias for each single observer, the effect on the final dataset, if any, should be substantially smoothed out as each user is presented with a differently ordered sequence of images. The presence of the visual anchors should also contribute to a reduction of this bias as the user has static references to compare the images against.

Technically, the web interface consists in a set of PHP pages and JavaScript scripts which record the answers of the user and store them on a PostgreSQL database.

The images used in the experiment were taken from the DERMOFIT database[11] and were captured with a fixed camera set-up (fixed position and controlled lighting) through a Canon EOS 350 DSLR equipped with a SIGMA 70mm f2.8 macro lens. No post-processing was performed on the images, apart from cropping them to a standard size of 600x450 pixels. Cropped images present the lesion in the middle surrounded by a patch of normal skin.

As previously mentioned, the interface (see Figure 2) differs from the one used in previous experiments[7] only through the introduction of the visual references which can be seen at each end of the sliders and in the middle. The design of the visual anchor points was quite important for guaranteeing accurate experiments. For this reason each visual anchor has been validated by the Department of Dermatology of the University of Edinburgh

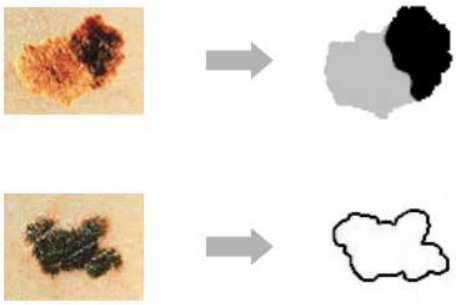

Figure 3. Example of original images (left) and obtained anchors (right) for asymmetry (top) and border irregularity (bottom). Original images obtained from the American Academy of Dermatology.[14]

before being included in the web-interface.
The first design choice was that of using cartoon-like graphics, instead of real lesion images, in order to help the user focus on the properties under evaluation one at a time. The main risk of using real images would have been that of having the user evaluation affected by properties not under scrutiny but suggesting high similarity between the sample and one visual anchor. As reported in many studies of similarity perception,[12,13] colour is often one of the most influential properties when evaluating the likeness of different images. If real images were used as anchor points, people might have been misled to move one slider, say the one for asymmetry, towards one of the references only for a resemblance in colour between the image under assessment and the visual anchor. The stylised grey-scale endpoint images are less prone to this undesirable effect and careful attention has been paid to select images carrying as little information as possible about the properties not directly linked to the afferent slider. The only exceptions are the anchors for texture, where patches of real images needed to be used as creating artificial samples satisfactorily representative of real lesions would have been impossible.
Whenever possible the visual anchors were obtained through graphical elaboration of the examples provided by the American Academy of Dermatology on their web-page illustrating the ABC(D) rule.[14] This was the case for all the references given for *colour uniformity* and for the upper extrema for both *asymmetry* and *border regularity*. All the other visual references have been obtained from real images of the DERMOFIT database[11] after discussion with the team of dermatologists. Some examples showing the original images and the obtained anchors are reported in Figure 3.
The survey has been proposed to three different categories of people, ranked on their level of education in skin-lesion assessment. The wider class consisted in people with no medical training and the experiment could be considered as a simulation of a self-screening procedure conducted on a variety of skin lesions. The results from this group were expected to shed more light on how precisely people evaluate the key properties on which self-screening guidelines, in particular the ABC(D) rule, rely upon. The second group was that of dermatologists. Given the substantial prior knowledge dermatologists have about skin lesions, this was considered a control group to verify whether variability in the evaluation is mainly due to lack of prior knowledge of laypeople or rather to differences in personal perception and assessment. Finally a group of people with some dermatology-related knowledge (medical doctors with different specialisations, nurses, medicine/nursing students, etc.) was included as an intermediate level between the two extremes. At the time of writing, answers have been obtained from 33 laypeople, 4 dermatologists and 5 non-dermatologist medically-trained people.

3.2 Data Analysis

The first and probably most important part of the data analysis was dedicated to evaluate how the inclusion of visual anchors affected the answers given by the volunteers who took part in the experiment. The reason behind this experiment was, in fact, understanding whether the extremely high variance in the scoring reported by Laskaris[7] was due to the lack of standardisation of the extrema of the scoring scale or rather to a more intrinsic variability in the answers linked to the subjectivity of the evaluation process. The comparison between answers obtained for a sample image before and after the inclusion of the visual anchors is reported in Figure 4.
Only the 22 images shown both in this experiment and in previous ones were considered in this part. For each of them, statistical testing has been performed to verify if the inclusion of the visual references resulted in a

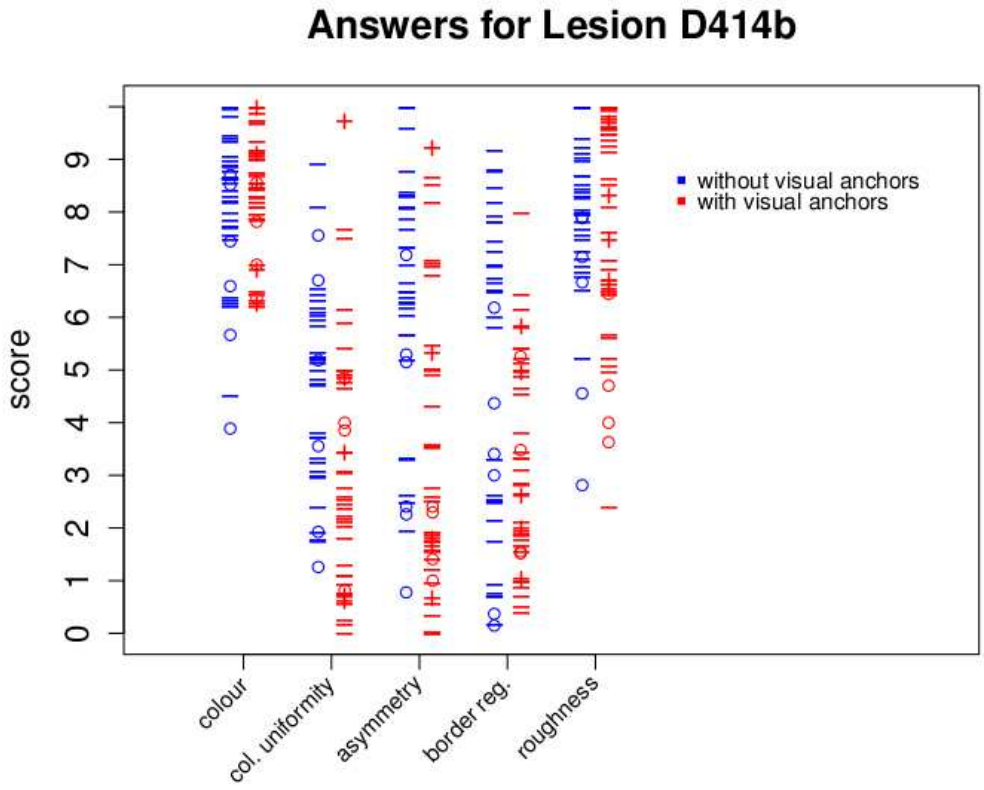

Figure 4. Impact of visual anchors on answers for lesion D414b. Marks encoding level of skin-related knowledge: '-' laypeople, '+' medically-trained people, '∘' dermatologists

statistically significant change in the variance of the answers. It is important to underline that only the variance of the measurements is comparable between the two sets of data, while any observed change in mean does not carry any useful information. This is due to the fact that, regardless the careful selection procedure, any choice of visual anchors is somehow arbitrary, especially for the central reference. It is reasonable to assume, hence, that a different set of visual references would result in a shift in mean, while the spread around this mean should remain substantially stable. A notable exception are those skin-lesion images whose average scoring for some of the properties is very near to the extrema of the scale. Since the score scale is bounded to the 0–10 range, in fact, the closer the score gets to the maximum (or minimum), the lower the variance tends to become as an effect of the upper (or lower) bound.

In the first place each of the 22 images was considered separately and the significance of the observed change in variance was tested. Given the specific characteristics of the scoring system (e.g. the limited 0–10 scoring range) and of the collected answers (such as small dimension of the samples, heavy-tails and skewness of their distribution) the data could not be considered to be distributed according to a Gaussian. The Bartlett's test[15] was hence inappropriate for the analysis due to its assumption of Normality of the data. As an alternative, the Brown–Forsythe[16] test was used. The Brown–Forsythe test is a variation of the Levene's[17] test in which the median is used in place of the mean of the sample. This difference makes the test more robust in cases where the data under analysis show a highly skewed distribution. Given the aforementioned bounded score scale, skewed distributions are often observed and hence this test is more appropriate. Specifically, the computed test statistic is

$$\frac{(N-k)}{(k-1)} \frac{\sum_{i=1}^{k} n_i (z_{i\cdot} - z_{\cdot\cdot})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (z_{ij} - z_{i\cdot})^2} \sim F_{k-1,N-k}$$

Table 1. Summary of the results of the Brown–Forsythe tests (at 95%) on changes in the variance after including the visual anchor points. Each column reports the number of significant changes for each property over the 22 test images.

	Significant Changes	**Reduction**	**Increase**
Colour	-	-	-
Colour Uniformity	5/22	3/22	2/22
Asymmetry	2/22	2/22	-
Border Regularity	10/22	10/22	-
Texture Roughness	6/22	2/22	4/22

where

$$z_{ij} = |y_{ij} - \tilde{y}_{i\cdot}|$$
$$z_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} z_{ij}$$
$$z_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} z_{ij}$$

and N is the total number of samples, k is the number of groups, n_i is the number of samples in group i, y_{ij} is the value of the j-th sample of the i-th group (in our case $i = \{1, 2\}$ representing the answers before and after the introduction of the visual anchors), $\tilde{y}_{i\cdot}$ is the median of the i-th group, $z_{\cdot\cdot}$ is the mean of all z_{ij}, $z_{i\cdot}$ is the mean of z_{ij} for elements of group i and $F_{k-1,N-k}$ represents the F distribution with $k-1$ and $N-k$ degrees of freedom.

The tests have been run at a 95% confidence level and the results are presented in table 1, where the number of the significant changes in variance is reported along with their direction (increase/reduction). While it is clear that the introduction of the visual anchors had no effect whatsoever on the variance of the answers obtained for the *colour* of the lesion, the other results cannot be interpreted without further analysis. The reason behind this necessity lies in the fact that since the gathered scores belong to the 0–10 real interval, they should be modelled as censored distributions, with censoring taking place both on the lower and on the upper side. What happens in practice is that as the mean approaches one of the extreme values, let us consider the lower bound 0 as an example, the data will progressively show less variability since no value lower than 0 is allowed. This shrinkage of the distribution is actually artificially induced by the bounded scale and for this reason all the variances obtained for values near the extremes are to be considered unreliable. While the statistical tests used cannot cope effectively with it, this situation should not be overlooked as an observed reduction in the variance might actually be the effect of a shift in the mean of the distribution to the region of one of the extreme values. As it turns out this is often the case. Table 2 is a more detailed version of table 1. For each statistically significant change detected by the Brown–Forsythe tests, the values of the variance and of the median before and after the inclusion of the visual anchors are reported. As it can be seen, most of the cases of statistically significant changes in the variance are actually associated with a shift of the median of the distribution (considered instead of the mean given the small dimension of the samples) towards one of the extremes of the 0–10 range. Finding a fixed value of the median above or below which the results of the test can be considered reliable is not easy, but if we assume the interval 2–8 to be a safe guess (having 20% of the possible 0–10 values on either side) we see that only 5 of the cases reported in table 2 have the median in this interval both before and after the inclusion of the visual anchors: one increment in variance for *colour uniformity* (lesion P206c), two reductions in variance for *border regularity* (lesions D414b and D578a) and two increments for *roughness of texture* (lesions P206b and P337a). Two border-line cases are represented by the significant increment in variance for the *roughness of texture* of lesions D414b and D578a despite the shift of their medians towards the upper bound.

On the basis of these considerations, it is important to interpret the data presented in table 3 with extreme care. These data were obtained testing the statistical significance of the changes in the average variance for each of the five properties with the Mann–Whitney test. The Mann–Whitney test[18] is an extension of the Wilcoxon test[18] to cases where the samples have different sizes. In turn, the Wilcoxon test is a non-parametric test often used in place of the Welch's t-test[19] when the assumption of Gaussianity does not hold for the samples.

Table 2. Statistics of the scores obtained before and after the introduction of the visual anchors for the statistically significant changes detected by the Brown–Forsythe tests.

Colour Uniformity				
Lesion reference	$\sigma^2_{\mathbf{before}}$	$\sigma^2_{\mathbf{after}}$	**Median before**	**Median after**
D262b	2.840	0.980	2.407	0.500
D726	3.266	1.00	1.593	0.667
P206c	1.985	5.709	6.444	5.426
P337a	6.027	0.525	1.926	0.538
P446	2.365	5.277	8.074	7.500
Asymmetry				
Lesion reference	$\sigma^2_{\mathbf{before}}$	$\sigma^2_{\mathbf{after}}$	**Median before**	**Median after**
D262b	5.608	2.498	3.074	0.741
P337c	2.344	0.807	1.704	0.370
Border Regularity				
Lesion reference	$\sigma^2_{\mathbf{before}}$	$\sigma^2_{\mathbf{after}}$	**Median before**	**Median after**
D262b	5.133	2.257	2.407	0.834
D270	6.620	1.723	2.815	1.315
D414b	8.849	3.746	6.000	2.982
D578a	9.402	4.667	6.593	4.519
D726	2.984	1.890	2.037	1.019
P257	7.762	2.928	2.556	1.241
P306a	8.344	2.064	2.741	1.352
P337a	1.638	0.180	1.037	0.204
P337c	1.714	0.450	1.370	0.241
P337e	5.781	2.599	4.000	0.889
Texture Roughness				
Lesion reference	$\sigma^2_{\mathbf{before}}$	$\sigma^2_{\mathbf{after}}$	**Median before**	**Median after**
D262b	4.820	1.134	2.519	0.908
D414b	2.123	4.443	7.963	8.222
D578a	1.904	5.041	8.296	8.482
D726	3.972	1.718	2.630	1.037
P206b	3.737	6.558	6.926	6.834
P337a	3.798	7.601	7.667	7.074

Table 3. Results of the Mann–Whitney (one-sided) tests on the change of the average variance after the inclusion of the visual anchors. The alternative hypothesis is presented in column H_a.

	$\sigma^2_{\mathbf{before}}$	$\sigma^2_{\mathbf{after}}$	$\mathbf{H_a}$	**p-value**
Colour	2.230	2.386	$\sigma^2_{before} < \sigma^2_{after}$	0.3086
Colour Uniformity	3.892	4.190	$\sigma^2_{before} < \sigma^2_{after}$	0.2317
Asymmetry	5.030	5.073	$\sigma^2_{before} < \sigma^2_{after}$	0.3507
Border Regularity	5.600	3.417	$\sigma^2_{before} > \sigma^2_{after}$	0.0004
Texture Roughness	4.955	5.840	$\sigma^2_{before} < \sigma^2_{after}$	0.0542

Table 4. Variance for each of the properties subdivided for level of skin-related education (n: number of participants). The values within the square brackets represent the minimum, median and maximum variance observed for the specific property on the whole set of images.

	Laypeople (n=33)	**Medically-trained** (n=5)	**Dermatologists** (n=4)
Colour	[0.99 – 2.33 – 3.90]	[0.26 – 2.21 – 8.47]	[0.01 – 1.29 – 7.00]
Asymmetry	[0.98 – 5.20 – 8.92]	[0.06 – 4.01 – 12.51]	[0.01 – 1.16 – 18.53]
Border Regularity	[0.17 – 3.75 – 7.06]	[0.07 – 2.34 – 10.02]	[0.02 – 0.74 – 11.61]
Colour Uniformity	[0.61 – 4.46 – 7.26]	[0.13 – 3.12 – 13.98]	[0.04 – 3.19 – 11.45]
Roughness of Texture	[1.33 – 5.60 – 11.76]	[0.08 – 3.78 – 14.63]	[0.16 – 1.74 – 7.19]

As pointed out before, there is no statistical means of deciding which of the cases should be considered and which should be ignored because their median is too close to one of the extreme values of the scoring scale. For this reason all the data have been included in the test, but the results must be considered with the *caveat* that the reliability of reductions or increments cannot be guaranteed and each specific case must be separately evaluated. In particular, the extremely high significance of the reduction of the average variance for *border regularity* could arguably be considered a wrong estimate as table 2 shows clearly that most of the cases of reduction in the variance of this property are obtained when the median is near either 0 or 10. The other significant change (nearly at 95%) is the increase in the average variance of *texture roughness* observed despite the fact that 4 out of the 6 lesions for which the change is significant had the median of the recorded score moved towards one of the extremes. Considering this, it can probably be concluded that such result is reliable. The other three properties (*colour*, *colour uniformity* and *asymmetry*) do not show any significant change and hence there is no reason to question the soundness of the associated tests. However, even if most of the changes in variance were significant, they would be increases in variance. This would suggest that the introduction of the visual anchors makes the consistency even worse.

4. CONCLUSIONS

Overall, two conclusions can be drawn. First of all the inclusion of the visual anchors did not have any considerable impact on the variability of the answers. The only statistically significant result seems to be an increase in the variance measured for the evaluation of the *roughness of texture*, while the reliability of the figures obtained for the *border regularity* is debatable.

These results are quite important as they prove that the variability observed in the evaluation of skin lesions obtained following qualitative guidelines like the ABC(D) rule is not mainly due to a subjective interpretation of the concepts on which the guideline is based (e.g. regular/irregular, symmetric/asymmetric, ...) since the inclusion of visual cues did not reduce the observed variance. If the variability could be really ascribed to the intrinsic subjectivity of the assessment, as the experimental results seem to suggest, the usefulness of guidelines like the ABC(D) rule would be under serious question. Secondly, even after the inclusion of the visual anchors, the value of the variance is quite high. The obtained standard deviations, in fact, range from a minimum of 1.545 for *colour* to a maximum of 2.417 for *roughness of texture*, which are quite high when considering that the scoring values range between 0 and 10.

The analysis of the difference between the answers given by people with different level of skin-related knowledge could not be performed as precisely as hypothesised in the first place. This was due to the fact that 4 dermatologists and 5 medically-trained people are insufficient to get any meaningful estimate of the variance within these groups. It can be reported, though, that the range in which the answers of these two groups are observed is comparable to that of laypeople, suggesting that the subjectivity of the evaluation dominates on other elements such as prior knowledge derived by education or experience. On the other hand, data presented in table 4 show an apparent reduction both in the minimum variance and in the median variance linked to the increase of skin-related knowledge. This evidence would suggest the presence of strong outliers in the dermatologists' data influencing the computation of the range. Given that these final observations are based on a limited amount of data, a wider data collection campaign will be required in the future to understand the importance of education and experience in the assessment of qualitative properties like those considered in this research.

ACKNOWLEDGMENTS

We thank the Wellcome Trust for funding this project.

REFERENCES

[1] Friedman, R. J., Rigel, D. S., and Kopf, A. W., "Early detection of malignant melanoma: the role of physicians examination and self examination of the skin," *CA Cancer Journal for Clinicians* **35**, 130–151 (1985).

[2] Weinstock, M. A., "ABCD, ABCDE, and ABCCCDEEEEFNUE," *Arch. Dermatol.* **142**, 528 (2006).

[3] Brändström, R., Hedblad, M., Krakau, I., and Ullén, H., "Laypersons' perceptual discrimination of pigmented skin lesions," *Journal of the American Academy of Dermatology* **46**, 667–673 (May 2002).

[4] Gunasti, S., Mulayim, M. K., Fettahlioglu, B., Yucel, A., Burgut, R., Sertdemir, Y., and Aksungur, V. L., "Interrater agreement in rating of pigmented skin lesions for border irregularity," *Melanoma research* **18**, 284–288 (2008).

[5] Meyer, L. J., Piepkorn, M., Goldgar, D. E., Lewis, C. M., Cannon-Albright, L. A., Zone, J. J., and Skolnick, M. H., "Interobserver concordance in discriminating clinical atypia of melanocytic nevi, and correlations with histologic atypia," *Journal of the American Academy of Dermatology* **34**, 618–625 (April 1996).

[6] Reetz Müller, K., Rangel Bonamigo, R., Antoniolli Crestani, T., Chiaradia, G., and Widholzer Rey, M. C., "Evaluation of patients' learning about the ABCD rule: a randomized study in southern Brazil," *Anais Brasileiros de Dermatologia* **84** (November/December 2009).

[7] Laskaris, N., *Fuzzy Description of Skin Lesion Images*, Master's thesis, School of Informatics – University of Edinburgh (2009).

[8] Laskaris, N., Ballerini, L., Fisher, R. B., Alridge, B., and Rees, J., "Fuzzy description of skin lesions," in [*Medical Imaging 2010: Image Perception, Observer Performance, and Technology Assessment*], Manning, D. J. and Abbey, C. K., eds., *Proceedings of the SPIE* **7627** (2010).

[9] Zanotto, M., *Visual Description of Skin Lesions*, Master's thesis, School of Informatics – University of Edinburgh (2010).

[10] Aldridge, R. B., Zanotto, M., Ballerini, L., Fisher, R. B., and Rees, J. L., "Novice identification of melanoma: not quite as straightforward as the ABCDs," *Acta Dermato-Venereologica* . In press.

[11] "http://homepages.inf.ed.ac.uk/rbf/dermofit/." Accessed on 16/01/2011.

[12] Mojsilović, A., Kovačević, J., Hu, J., Sarfranek, R. J., and Ganapathy, S. K., "Matching and retrieval based on vocabulary and grammar of color patterns," *IEEE Transactions on Image Processing* **9**, 38–54 (January 2000).

[13] Rogowitz, B. E., Frese, T., Smith, J. R., Bouman, C. A., and Kalin, E., "Perceptual image similarity experiments," in [*Human Vision and Electronic Imaging III*], Rogowitz, B. E. and N., P. T., eds., *Proceedings of the SPIE, 3299* (January 1998).

[14] "http://www.skincarephysicians.com/skincancernet/melanoma.html." Accessed on 16/01/2011.

[15] Bartlett, M. S., "Properties of sufficiency and statistical tests," in [*Proceedings of the Royal Statistical Society*], *A* **160**, 268–282 (1937).

[16] Brown, M. B. and Forsythe, A. B., "Robust tests for equality of variances," *Journal of the American Statistical Association* **69**, 364–367 (1974).

[17] Levene, H., "Robust tests for equality of variances," in [*Contributions to probability and statistics: essays in honor of Harold Hotelling*], Olkin, I., ed., 278–292, Stanford University Press, Stanford, CA (1960).

[18] Corder, G. W. and Foreman, D. I., [*Nonparametric Statistics for non-statisticians*], John Wiley and Sons (2009).

[19] Welch, B. L., "The generalization of student's problem when several different population variances are involved," *Biometrika* **34**(1–2), 28–35 (1947).